

Refinements of the Eigenstate Thermalization Hypothesis under Local Rotational Invariance via Free Probability

Elisa Vallini^{1,*}, Laura Foini², and Silvia Pappalardi¹

¹*Institut für Theoretische Physik, Universität zu Köln, Zùlpicher Straße 77, 50937, Köln, Germany*

²*Institut de Physique Théorique, CNRS, CEA, Université Paris-Saclay, 91191, Gif-sur-Yvette, France*

 (Received 5 December 2025; revised 1 July 2026; accepted 4 August 2026; published 13 October 2026)

The eigenstate thermalization hypothesis (ETH) was developed as a framework for understanding how the principles of statistical mechanics emerge in the long-time limit of isolated quantum many-body systems. Since then, ETH has shifted the attention toward the study of matrix elements of physical observables in the energy eigenbasis. In this work, we revisit recent developments leading to the formulation of full ETH, a generalization of the original ETH ansatz that accounts for multi-point correlation functions. Using tools from free probability, we explore the implications of local rotational invariance, a property that emerges from the statistical invariance of observables under random basis transformations induced by small perturbations of the Hamiltonian. This approach allows us to make quantitative predictions and derive an analytical characterization of subleading corrections to matrix-element correlations, thereby refining the ETH ansatz. Moreover, our analysis links the statistical properties of matrix elements under random basis changes to the empirical averages over energy windows that are usually considered when dealing with a single instance of the ensemble. We validate our analytical predictions through comparison with numerical simulations in non-integrable Floquet systems.

DOI: [10.1103/d1cq-gj9p](https://doi.org/10.1103/d1cq-gj9p)

I. INTRODUCTION

The eigenstate thermalization hypothesis (ETH) was introduced as a natural extension of ideas from quantum chaos, typicality, and random matrix theory (RMT) to explain thermalization in isolated quantum systems [1–7]. In the quantum chaos community, it is common to model deterministic systems as effectively random. In this spirit, the Bohigas-Giannoni-Schmit conjecture established that quantum chaotic systems exhibit eigenvalue statistics consistent with RMT [8]. Beyond spectral properties, quantum chaos also manifests in the statistics of eigenstates. According to Berry’s random wave conjecture, eigenfunctions of chaotic systems $H|E_i\rangle = E_i|E_i\rangle$ behave as superpositions of plane waves with Gaussian-distributed amplitudes and random phases [9,10].

Building on these foundations, ETH focuses on the matrix elements of few-body observables in the energy eigenbasis, $A_{ij} = \langle E_i|A|E_j\rangle$. It has been proposed that these matrix elements can be regarded as random variables: the average of the diagonal elements varies smoothly

with energy and reproduces thermal expectation values, while the off-diagonal elements have zero mean and an energy-dependent variance [3,4]. These small fluctuations can be viewed as arising from a fictitious ensemble of infinitesimally perturbed systems, effectively introducing randomness through the arbitrary mixing of nearby eigenstates, without altering the relevant physical properties at the scales of interest in statistical mechanics [5]. ETH connects the statistical properties of the matrix elements to the thermal behavior of A , thereby explaining thermalization in isolated chaotic systems. Randomness emerges as the underlying mechanism: subject to the constraint of preserving a minimal structure in energy, the chaotic eigenstates are otherwise as random as possible, encoding the microcanonical ensemble [4,11–16].

Recent developments have extended the standard ETH framework by studying higher-order correlations among matrix elements [17]. These correlations are now known to play a crucial role in determining quantities such as out-of-time-order correlators (OTOCs) and have been investigated in different contexts, ranging from condensed matter [18–23] to high energy physics [24–31]. The *full ETH* [17] states the following expression, for products with distinct indices $i_1 \neq \dots \neq i_n$:

$$\overline{A_{i_1 i_2} \dots A_{i_n i_1}} = D_{E_+}^{1-n} F_{E_+}^{(n)}(\omega_1, \dots, \omega_n) + \mathcal{O}(D_{E_+}^{2-n}), \quad (1a)$$

*Contact author: evallini@uni-koeln.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

where $F^{(n)}$ are some smooth functions of the energy differences $\omega_\alpha = E_{i_\alpha} - E_{i_{\alpha+1}}$ and of the mean energy density $e_+ = E_+/L = (E_{i_1} + \dots + E_{i_n})/nL$, with L the system size. They are commonly known as *ETH smooth functions* and are related to the Fourier transforms of higher-order connected thermal correlation functions. $D_{E_+} = e^{S(E_+)}$ is the density of states at the average energy, with S macroscopic entropy. When indices are repeated, the average of the products factorizes, at the leading order in the density of states D_{E_+} , as

$$\overline{A_{i_1 i_2} \dots A_{i_{m-1} i_m} A_{i_1 i_{m+1}} \dots A_{i_n i_1}} \approx \overline{A_{i_1 i_2} \dots A_{i_{m-1} i_m}} \overline{A_{i_1 i_{m+1}} \dots A_{i_n i_1}}. \quad (1b)$$

Remarkably, it has been recognized that this ansatz can be reformulated within the mathematical framework of *free probability theory* [32,33]. The latter generalizes classical probability to non-commuting variables by replacing the notion of statistical independence with *freeness* [34–36]. Specifically, it introduces powerful tools for understanding the statistical properties of large matrices, such as those that arise in chaotic quantum systems. Recasting ETH in this language has led to several conceptual developments [20,37–41], including the intriguing proposal that, in ergodic systems, local observables at different times become freely independent in the long-time limit [42–47].

In this work we further develop the *principle of local rotational invariance*, which originally motivated the formulation of the full ETH [48]. On the one hand, we clarify in detail how the full ETH emerges from local rotational invariance; on the other hand, pursuing this goal also provides a quantitative characterization of the subleading corrections to Eq. (1), which were only partially discussed in [17].

Rotational invariance can be realized at different levels. In its simplest implementation, basis transformations are taken to be global: the probability distribution of the Hamiltonian remains unchanged under arbitrary unitary transformations [49], leading to completely random eigenvectors uniformly distributed on the unit sphere [50,51]. Consequently, the statistical properties of A remain invariant under random global unitary rotations. In realistic Hamiltonians, however, randomness emerges only locally in energy: eigenstates exhibit random-like fluctuations on small energy scales while preserving smooth variations and physical structure across the spectrum [52–57]. This observation motivated refined RMT models, such as banded random matrix ensembles, where randomness is restricted to nearby energy levels [58,59]. Within these models, the statistical properties of observables remain invariant only under *local* unitary transformations, providing a more faithful statistical description of chaotic many-body systems. The transition from global to local

rotational invariance can thus be viewed as a refinement of the ETH paradigm [17]: the global case represents the featureless random limit, while the local version incorporates energy structure. This hierarchical picture provides a controlled approach to study the internal structure of ETH and to explicitly evaluate its first subleading corrections, which naturally emerge in terms of *local free cumulants* [60], as we aim to show in this work. These corrections are compatible with the postulates of structured random matrices introduced in [60] beyond the leading order.

A. Motivation and main results

Before entering the technical development, we summarize here the physical idea behind our approach and our main results.

a. Local rotational invariance and the ensemble U_{loc} . Our starting point is the physical observation that the observable statistical properties that are relevant for thermalization should not change if we slightly perturb the Hamiltonian in a way that only reshuffles eigenstates that are close in energy. Indeed, mesoscopic physics is expected to be insensitive to such small perturbations, which mix only nearby eigenvectors while leaving the smooth, coarse-grained structure of the spectrum intact. This expectation is encoded in the *principle of local rotational invariance* [48]: an observable A in the energy eigenbase shares its statistical properties with a version $A^U = U_{\text{loc}}^\dagger A U_{\text{loc}}$ rotated by a random unitary that acts *only locally* in energy,

$$A \sim A^U, \quad U_{\text{loc}} = \bigoplus_k U^{(I_k)}. \quad (2)$$

Each $U^{(I_k)}$ is an independent Haar-random unitary that mixes only the states inside a small energy window of width Δ , centered on a mesoscopic energy E_i . Such a window is described by a function $\delta_\Delta(E_i - E_\alpha)$, either a sharp box or a smoothed profile peaked around E_i , as illustrated in Fig. 1(a), so that the number of states it contains is $D_i = \sum_\alpha \delta_\Delta(E_i - E_\alpha) \simeq e^{S(E_i)}$. Averaging over U_{loc} reshuffles the matrix elements of A only locally in energy, while preserving a hierarchy of correlations.

b. Main results. In this work, we compute averages of products of matrix elements, such as those appearing in Eq. (1). Local rotational invariance allows these quantities to be computed as ensemble averages over the corresponding unitaries (2). Carrying out this calculation yields the full ETH ansatz, showing that local rotational invariance provides the underlying principle governing the structure of the correlations encoded in ETH. The explicit evaluation further leads to the following results:

- (1) the definition of the *smooth ETH functions in terms of global matrix quantities*, which make the energy dependence of matrix elements explicit;

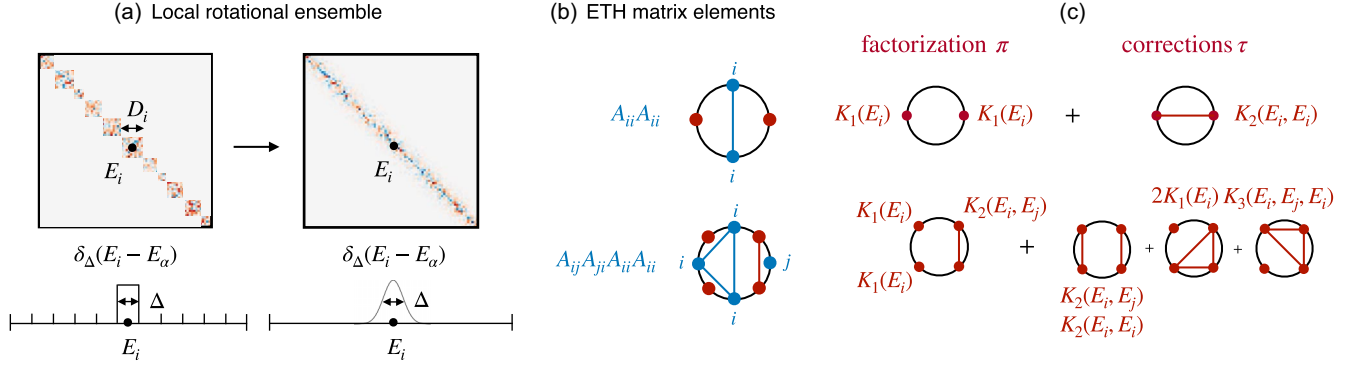


FIG. 1. *Pictorial summary.* (a) The local rotational ensemble U_{loc} : a random unitary with block structure (left) or with a smoothed/banded structure in energy (right). Only the D_i states within an energy window of width Δ around each mesoscopic energy E_i are mixed by a matrix of the ensemble. This window is described by the function $\delta_\Delta(E_i - E_\alpha)$, which may take the form of a sharp box or a smooth profile. (b) Products of matrix elements, such as $A_{ii}A_{ii}$ and $A_{ij}A_{ji}A_{ii}A_{ii}$, are represented by blue diagrams on n points, where repeated indices are connected by blue lines. This uniquely fixes π , identified by the associated diagram on the red edges. (c) Leading factorization and subleading corrections of correlations with repeated indices [cf. Eq. (9)]. The leading contribution is determined by π and given by the product over its blocks of smooth functions evaluated at the energies labeling the nodes. Subleading corrections are obtained from diagrams τ formed by merging two blocks of π . Top row: $k = 2$ factorization in Eq. (5). Bottom row: $k = 4$ factorization in Eq. (8).

- (2) the definition of the *empirical averages* entering the ETH ansatz. While the derivation relies on ensemble averages over random rotations, these averages ultimately admit a precise formulation in terms of matrix elements of a single deterministic operator.
- (3) the identification of *subleading corrections* to the ETH factorization;

These are our main results.

The simplest example: the standard ETH. In this paragraph, we illustrate these results on the simplest examples at first and second order, which correspond to the *standard ETH ansatz* [4,7].

When the indices are distinct, the average over the local rotational invariant ensemble gives *smooth functions of the energies* (result 1.) as

$$\overline{A_{ii}^U}(E_i) = \text{Tr}(\hat{\sigma}_{E_i} A) \equiv K_1(E_i), \quad (3a)$$

$$\begin{aligned} \overline{A_{ij}^U A_{ji}^U}(E_i, E_j) &\simeq \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_j} A) - [\text{Tr}(\hat{\sigma}_{E_i} A)]^2 \sigma(E_i - E_j) \\ &\equiv K_2(E_i, E_j), \end{aligned} \quad (3b)$$

where $\hat{\sigma}_E = \sum_\alpha \frac{\delta_\Delta(E - E_\alpha)}{D_E} |\alpha\rangle\langle\alpha|$ is a normalized microcanonical projector peaked around the energy E , and $\sigma(E_i - E_j) = \delta_\Delta(E_i - E_j)/D_{E_i}$. In this way we obtain energy-resolved moments of A , i.e. traces projected onto the energy windows, which define the smooth functions of the energy $K_1(E) = F_e^{(1)}$ and $K_2(E_1, E_2) = F_{e^+}^{(2)}(\omega)/D_{E_+}$ [using the notation in Eq. (1)].

Furthermore, we note that, for $i \neq j$, we can rewrite Eq. (3b) as

$$\overline{A_{ij}^U A_{ji}^U} \simeq \sum_{\alpha \neq \beta} A_{\alpha\beta} \frac{\delta_\Delta(E_i - E_\alpha)}{D_i} A_{\beta\alpha} \frac{\delta_\Delta(E_j - E_\beta)}{D_j}. \quad (4)$$

As we discuss below, this shows that ETH averages can be interpreted as averages over small energy windows of a single realization of the operator, the *empirical averages* entering the ETH ansatz (result 2.).

Finally, when some indices are repeated, the average of the matrix elements *factorizes* into products of these smooth functions, plus a *subleading correction* (result 3.); the simplest instance is the diagonal two-point function

$$\overline{A_{ii}^U A_{ii}^U}(E_i) \simeq \overline{A_{ii}^U}(E_i) \overline{A_{ii}^U}(E_i) + \overline{A_{ij}^U A_{ji}^U}(E_i, E_i), \quad (5)$$

where the first term is the factorization and the second, of relative order $1/D$, is the subleading correction. As we explain better below, the latter encodes a nontrivial *gluing* of leading contributions, mechanism that generates higher-order correlations, although constrained to equal-energy sectors. Equation (5) is the well-known statement that the variance of the diagonal matrix elements is fixed by the off-diagonal ones at zero frequency [7]. Indeed, the subleading contribution is already present in the standard ETH, but missing in the usual higher-order formulations. One of the goals of this paper is to extend these results to arbitrary orders, thereby filling a structural gap in the full ETH framework.

The general result. We now show how the results above hold at arbitrary order.

For an all-distinct cyclic product (a simple loop), the average is a single smooth function of the energies,

$$\overline{A_{i_1 i_2}^U A_{i_2 i_3}^U \cdots A_{i_n i_1}^U} \simeq K_n(E_{i_1}, \dots, E_{i_n}), \quad (6)$$

which extends Eqs. (3a)–(3b) to arbitrary n . We will be able to write the smooth functions K_n in terms of moments of A resolved in energies by microcanonical projectors $\hat{\sigma}_E$, and to show that these expressions are *local free cumulants* [60], directly connected to operator valued free probability. These provide a natural definition of the ETH smooth functions, $F_{e_+}^{(n)}() = \lim_{N \rightarrow \infty} D_{E_+}^{n-1} K_n()$ in terms of global matrix quantities.

A key outcome is that the ensemble averages coincide with *empirical averages* of a single, fixed operator over small energy windows. Using $A \sim A^U$ and self-averaging, the diagrams with distinct indices read

$$\begin{aligned} \overline{A_{i_1 i_2}^U \cdots A_{i_n i_1}^U} &\simeq \sum_{\alpha_1 \neq \cdots \neq \alpha_n} A_{\alpha_1 \alpha_2} \frac{\delta_\Delta(E_{i_1} - E_{\alpha_1})}{D_{i_1}} \cdots \\ &\cdots A_{\alpha_n \alpha_1} \frac{\delta_\Delta(E_{i_n} - E_{\alpha_n})}{D_{i_n}}, \end{aligned} \quad (7)$$

thus generalizing Eq. (4). The right-hand side involves only the matrix elements of the fixed operator A , summed over normalized energy windows around $E_{i_1}, \dots, E_{i_n}$: the ensemble randomness is replaced by a *smoothing over small energy windows* of a single deterministic matrix. Nicely, this provides the standard recipe used to extract ETH functions numerically, and it extends to repeated indices by smoothing only over the distinct ones.

The factorization at repeated indices, together with its subleading corrections, also admits a simple interpretation as a generalization of the mechanism already highlighted in Eq. (5). At the fourth order, for instance,

$$\begin{aligned} \overline{A_{ij}^U A_{ji}^U A_{ii}^U A_{ii}^U} &\simeq \overline{A_{ij}^U A_{ji}^U}(E_i, E_j) \overline{A_{ii}^U}^2(E_i) \\ &+ \overline{A_{ij}^U A_{ji}^U}(E_i, E_j) \overline{A_{ii}^U A_{ii}^U}(E_i, E_i) \\ &+ 2 \overline{A_{ii}^U}(E_i) \overline{A_{ij}^U A_{jk}^U A_{ki}^U}(E_i, E_j, E_i), \end{aligned} \quad (8)$$

that is, a leading factorization followed by subleading corrections built from products with one more distinct index, evaluated at equal mesoscopic energies.

This mechanism can be understood pictorially from Figs. 1(b)–1(c). Firstly, we associate to each product of matrix elements a diagram π , as illustrated in Fig. 1(b) for the examples of Eqs. (5) and (8). The matrix elements are represented by blue dots, joined by a blue line whenever their indices coincide, while the diagram π is read off from the red edges.

We focus on products for which this diagram is non-crossing and show that the average reads

$$\overline{[A_{i_1 i_2}^U \cdots A_{i_n i_1}^U]_\pi} \simeq \underbrace{K_\pi(\{E\})}_{\text{factorization}} + \sum_{\tau} \underbrace{K_\tau(\{E\})}_{\text{subleading}}. \quad (9)$$

The leading term K_π is read directly from the diagram π [Fig. 1(c), left]: it is given by the product of smooth functions K_n , with one factor for each block of π , evaluated at the energies associated with the dots in that block. The subleading corrections arise from diagrams τ obtained by merging two blocks of π [Fig. 1(c), right]; each K_τ is again a product of smooth functions. For instance, at $k = 2$ [top row of Figs. 1(b)–1(c)] the diagram of $A_{ii}^U A_{ii}^U$ has two blocks, yielding the factorization $[K_1(E_i)]^2 = \overline{A_{ii}^U}^2(E_i)$. Merging them gives the single correction $K_2(E_i, E_i) = \overline{A_{ij}^U A_{ji}^U}(E_i, E_i)$, which reproduces Eq. (5). In the body of the work, we will show that this procedure has a precise mathematical formulation and that π and τ are naturally organized on the lattice of non-crossing partitions, appearing in free probability.

Strategy of the paper. To build intuition, we begin with the simplest case of *global* rotational invariance, i.e. a single global Haar rotation acting on all matrix elements. This corresponds to taking a single block of size D , equal to the full Hilbert space dimension, in U_{loc} . This toy model already captures the factorization properties and the structure of the subleading corrections, but is blind to any energy dependence: in this case the local free cumulants reduce to the normalized free cumulants of the observable A , $K_n() = \kappa_n/D^{n-1}$, independent of energy. We then refine this construction to local rotational invariance, which restores energy dependence and, crucially, yields the smooth energy dependence of the ETH functions.

B. Paper's structure

The rest of the manuscript is organized as follows. In Sec. II we present in detail the idea of global and local rotational invariance; we briefly review Weingarten calculus for the calculation of averages of products of matrix elements, and key concepts of free probability. The next four sections contain the bulk of our new results:

- (i) Section III: we start by the simplest toy model of global rotational invariance; we derive a closed formula for the leading and subleading contributions; those can be read in terms of free cumulants on the lattice of non-crossing partitions.
- (ii) Section IV: we consider the second toy model with local rotational invariance, which allows us to improve the closed formula from the global case by incorporating the energy dependence; we translate the corresponding quantities from standard free cumulants to local free cumulants, related to operator-valued free probability.

- (iii) Section V: these findings lead us to define empirical averages over small energy windows for the fixed matrix A and to directly link the smooth functions of ETH and local free cumulants. In this way, we gain back ETH and the first subleading correction to the factorization of the diagrams.
- (iv) Section VI provides numerical tests conducted in a many-body non-integrable Floquet system, demonstrating strong agreement with our theoretical predictions.

In Sec. VII we conclude and discuss the implications of our work, with an outlook on future directions suggested by the principle of local rotational invariance and possible applications of the subleading corrections.

There are several appendices. In Appendix A we introduce additional concepts from free probability that complement the ones in the main text. We then discuss the detailed calculations through Weingarten calculus in Appendixes B and C, respectively for global and local rotational invariance. In Appendix D we link the results obtained from local rotational invariance to operator-valued free probability. An alternative approach to free probability, based on the Harish-Chandra–Itzykson–Zuber integral, is discussed in Appendix E, while Appendix F presents additional numerical analyses. In Appendix G we show how the subleading corrections discussed in this paper determine the late-time plateau of *thermal free cumulants* [32], providing a quantitative characterization of long-time freeness in chaotic dynamics.

II. RANDOM BASIS TRANSFORMATIONS

In this work we study the matrix elements of some observable A in a randomly rotated basis by studying correlations between the entries of $A^{\mathcal{U}} = \mathcal{U}^\dagger A \mathcal{U}$ with $\mathcal{U}$ a unitary matrix drawn from an ensemble that we will specify. The basic assumption is to look at a single matrix A in the basis of the Hamiltonian and assume that it is characterized by *global or local rotational invariance*. With this we mean that A shares the statistical properties with a globally or locally rotated version:

$$A \sim A^{\mathcal{U}} = \mathcal{U}^\dagger A \mathcal{U}, \quad (10)$$

where $\mathcal{U}$ can be a Haar distributed random unitary matrix, as the one in Fig. 2(a), or a banded random unitary characterized by a local structure in energy, such as the block diagonal matrix in Fig. 2(b) or its smoothed version in Fig. 2(c). Moments of A are clearly equal to those of $A^{\mathcal{U}}$ as they depend only on the spectral properties of the matrix. In the context of local rotational invariance we will encounter generalized moments, and we will assume that in the large size limit these moments of A equal those of $A^{\mathcal{U}}$, and that in particular these quantities are self-averaging, namely that a single realization of the matrix equals the result on average.

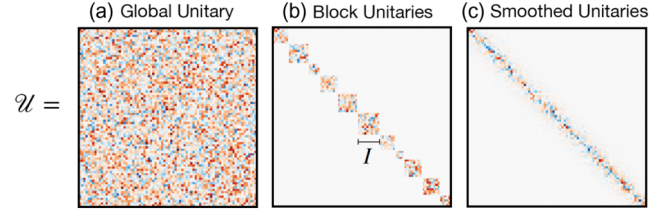


FIG. 2. Illustrative examples of the random matrix $\mathcal{U}$ in Eq. (10), employed to rotate the observable A . Different refined toy models for deriving ETH are shown. In (a) the matrix $\mathcal{U}$ is a global Haar unitary matrix, so that A exhibits consequently global rotational invariance. Local rotational invariance is achieved by incorporating a local structure, as in (b) and (c): in (b) the matrix $\mathcal{U}$ is composed of disjoint blocks of independent Haar unitaries, each acting only on a local interval of states. The last more refined toy model in (c) is a smoothed version of the previous one, see Sec. II for discussion.

Random matrix theory and free probability provide a framework for evaluating matrix elements' n -th moments $\overline{A_{i_1 j_1}^{\mathcal{U}} \cdots A_{i_n j_n}^{\mathcal{U}}}$, which are the interest of ETH. We explicitly keep the dependence on $\mathcal{U}$ to emphasize that we are averaging over an ensemble of random matrices, assuming $A \sim A^{\mathcal{U}}$. The empirical averages performed on the matrix elements of a fixed observable A will be discussed in Sec. V. We now define the different toy models that we will consider.

In the first toy model $\mathcal{U} = U$ is a structureless random unitary, and the observable A^U possesses global rotational invariance. Therefore, $A_{ij}^U = \sum_{\bar{i}, \bar{j}} U_{i\bar{i}}^* A_{\bar{i}\bar{j}} U_{\bar{j}j}$ with U sampled by the Haar ensemble. The matrix elements' n th moments follow as

$$\overline{A_{i_1 j_1}^U \cdots A_{i_n j_n}^U} = \sum_{\substack{\bar{i}_1 \cdots \bar{i}_n \\ \bar{j}_1 \cdots \bar{j}_n}} \overline{U_{i_1 \bar{i}_1}^* \cdots U_{i_n \bar{i}_n}^* U_{\bar{j}_1 j_1} \cdots U_{\bar{j}_n j_n}} A_{\bar{i}_1 \bar{j}_1} \cdots A_{\bar{i}_n \bar{j}_n}. \quad (11)$$

The statistical properties of the Haar ensemble are known exactly and can be expressed compactly using the Weingarten calculus, which follows from the Schur–Weyl duality and Schur's lemma, allowing unitary invariants to be written in terms of permutation operators as [61]

$$\overline{U_{i_1 \bar{i}_1}^* \cdots U_{i_n \bar{i}_n}^* U_{\bar{j}_1 j_1} \cdots U_{\bar{j}_n j_n}} = \sum_{\alpha, \beta \in S_n} \text{Wg}_{\alpha, \beta}(D) \delta_{i_{\alpha(1)}, \bar{j}_1} \cdots \delta_{i_{\alpha(n)}, \bar{j}_n} \delta_{\bar{i}_1, j_{\beta(1)}} \cdots \delta_{\bar{i}_n, j_{\beta(n)}}, \quad (12)$$

with $\text{Wg}_{\alpha, \beta}(D) = \text{Wg}(\alpha^{-1}\beta, D)$ Weingarten matrix for α, β permutations in the symmetric group of n elements S_n . D is the dimension of the matrix A i.e., the Hilbert space dimension. Further discussion of the Weingarten function will be given in Appendix B.

A different approach based on an explicit formula for the generating function of the matrix elements, the (low rank limit of the) Harish-Chandra-Itzykson-Zuber integral, is discussed in Appendix E.

The second toy model takes into account that A remains invariant under local basis transformations i.e. reshufflings of its elements restricted to local neighborhoods in energy. We consider $\mathcal{U} = U_{\text{loc}}$ to have local structure in the energy of the Hamiltonian, i.e. $H|E_i\rangle = E_i|E_i\rangle$. We divide the total energy range in M disjoint intervals, $I_1 \cdots I_M$. The size of each interval, in terms of energy, is denoted by Δ , and intuitively all states in the same block are associated with the same *mesoscopic* energy. Considering a generic interval I , whichever $i \in I$ can be taken as representative of the associated mesoscopic energy that we therefore label by the state, as E_i . The number of states of I , that we call D_i , is defined as

$$D_i \equiv \sum_{\alpha \in I} 1 = \sum_{\alpha} \delta_{\Delta}(E_i - E_{\alpha}) \simeq e^{S(E_i)}, \quad (13)$$

where the $\delta_{\Delta}(E_i - E_{\alpha})$ selects states within the block I to which i belongs, as a sharp box, and $e^{S(E_i)}$ is the density of states at energy E_i , with S the macroscopic entropy. One can therefore define a matrix $U^{(I)}$ for each interval I , as a block-sparse matrix with a single random unitary block of size $D_i \times D_i$, supported on the indices I . Globally, as depicted in Fig. 2(b),

$$U_{\text{loc}} = \bigoplus_{k=1}^M U^{(I_k)} \quad (14)$$

where $U^{(I_k)}$ are independent. As a consequence, the locally rotated matrix is

$$A_{ij}^U = (U_{\text{loc}}^\dagger A U_{\text{loc}})_{ij} = \sum_{\substack{\tilde{i} \in I \\ \tilde{j} \in J}} U_{\tilde{i}i}^{(I)*} A_{\tilde{i}\tilde{j}} U_{\tilde{j}j}^{(J)}, \quad (15)$$

where now $U^{(I)}$ and $U^{(J)}$ are independent Haar unitaries for the blocks I and J respectively: the rotation is made locally around the energies of the states i and j . The matrix elements' n th moments for locally rotated matrices are

$$\begin{aligned} & \overline{A_{i_1 j_1}^U \cdots A_{i_n j_n}^U} \\ &= \sum_{\substack{\tilde{i}_1 \in I_1 \cdots \tilde{i}_n \in I_n \\ \tilde{j}_1 \in J_1 \cdots \tilde{j}_n \in J_n}} \overline{U_{\tilde{i}_1 i_1}^{(I_1)*} \cdots U_{\tilde{i}_n i_n}^{(I_n)*} U_{\tilde{j}_1 j_1}^{(J_1)} \cdots U_{\tilde{j}_n j_n}^{(J_n)}} A_{\tilde{i}_1 \tilde{j}_1} \cdots A_{\tilde{i}_n \tilde{j}_n}. \end{aligned} \quad (16)$$

Differently from the case of global rotational invariance, one must account for the fact that the matrices are random unitaries only on the associated set of indices. Therefore, Eq. (12) cannot be applied directly; however, Weingarten calculus can still be used by appropriately coupling

matrices on different intervals. One clearly retrieves the global Haar case for a single interval $M = 1$ with a flat $\delta_{\Delta}() = 1$ in Eq. (13).

This minimal toy model, although convenient for analytical computations, exhibits intrinsic discontinuities, since indices i and j that are close in energy may still belong to distinct blocks I and J . Nevertheless, it has a natural counterpart that can be schematically related to the banded unitary shown in Fig. 2(c). Indeed, its formulation allows one to generalize the minimal model into a smoothed version by considering $\delta_{\Delta}()$ as peaked around E_i for each state i with a characteristic width Δ . Moving from the sharp to the smoothed function yields a refined version of the local rotational invariance model.

In the next sections, Eqs. (11) and (16) will be evaluated explicitly through Weingarten calculus, and products of matrix elements will be given in terms of free cumulants. Since these concepts are essential to explain our results, we briefly review in the next paragraph the basic elements of free probability theory.

A. Key concepts of free probability

Free probability is an extension of traditional probability theory to non-commuting variables, such as random matrices. Originating in planar field theory [62–64], free probability has recently played a role in many branches of many-body physics, such as quantum information theory [65], tensor networks [66–69], disordered systems [70–75], and gravity [76–80].

This section provides a concise introduction to the aspects of free probability theory required to understand the main results of this paper, focusing on its combinatorial formulation. Additional material is presented in Appendix A, where we provide a more standard mathematical introduction to free probability and a brief overview of operator-valued free probability. The latter extends free probability to variables with additional structure, such as block random matrices, and as we show below, some of the structures emerging from our analysis admit a natural interpretation within this framework.

For a more comprehensive treatment of free probability theory, see Refs. [81–83].

In Fig. 3 some examples of the main concepts are illustrated.

a. Partitions. In order to describe the combinatorics of free probability, it is useful to introduce partitions. A partition of a set of n elements is a collection of non-empty, disjoint subsets whose union equals the original set; we indicate it as a sequence of blocks B_{μ} , $\mu = \{B_1, \dots, B_{\#\mu}\}$, with $\#\mu$ the total number of blocks in the partition. Each partition is usually represented graphically by n points on a circle, ordered clockwise, with lines connecting the elements in the same block. Crossing partitions are the ones for which *blocks cross*, i.e., lines in the respective diagram intersect. Non-crossing partitions are all the others, are

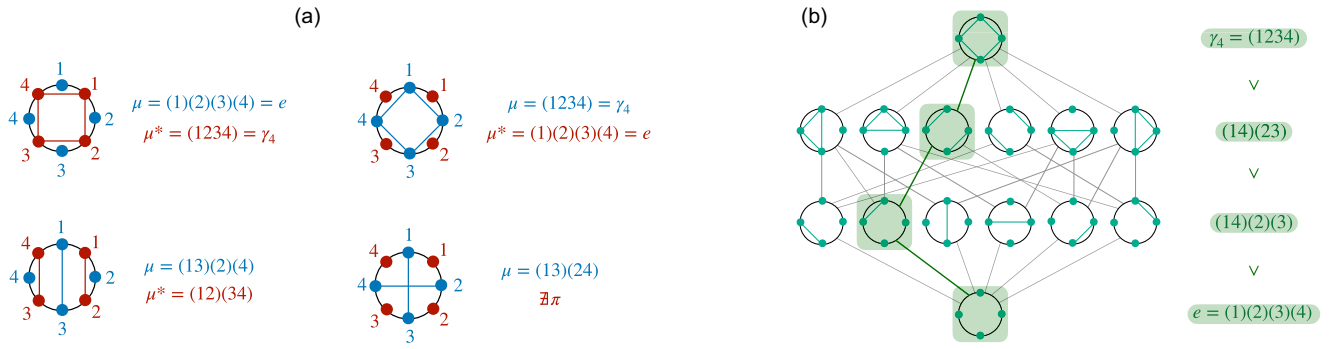


FIG. 3. (a) Graphical representation of some partitions of four elements (in blue) together with their respective Kreweras complements (in red). The unique crossing partition for $n = 4$ is shown in the lower right corner and has no Kreweras complement. (b) $NC(4)$ (sub)lattice of non-crossing partitions; lines between rows are connecting partitions at a unit distance. At the bottom, the discrete partition (finest), at the top the trivial one (coarsest). Correspondingly, $NC(4)$ can be seen as the geodesic interval $[e, \gamma_4]$, subset of the Cayley graph, for which permutations from one row to the other are related via a single transposition. Each set of permutations following a path from e to γ_4 defines a single geodesic in the interval; an explicit example is highlighted in green, and on the right.

denoted by $NC(n)$, and their number is given by the Catalan number $C_n = \binom{2n}{n}/(n+1)$. Two special partitions are the discrete partition e (n blocks of 1 element) and the trivial one γ_n (1 block of n elements).

b. Lattice. Partitions are partially ordered by refinement: for two partitions μ and ν , we write $\nu \leq \mu$ if every block of ν is contained in a block of μ . Under this order, the set of all partitions forms a lattice which can be visualized as a hierarchy of partitions arranged from the coarsest at the top, which is the trivial partition, to the finest at the bottom, which is the discrete partition. Not all partitions are comparable, which makes the refinement only a partial order: incomparable partitions appear on the same level of the lattice. When considering only non-crossing partitions, a similar structure can be drawn, forming a sublattice of the lattice of all partitions under the same refinement order. The sublattice of $NC(n)$ exhibits a top-bottom symmetry which is a special feature of non-crossing partitions and does not hold in the lattice of all partitions. The symmetry consists in exchanging the coarsest and finest partitions and reversing the order of all others; this leaves the lattice combinatorially unchanged. This inversion symmetry can be made precise, defining the Kreweras complement.

c. Dual partition or Kreweras complement. The Kreweras complement of a non-crossing partition μ is obtained diagrammatically by placing a second copy of the set $\{1, \dots, n\}$ around the same circle and taking the maximal non-crossing partition on the new points without crossing any lines of μ . Intuitively, μ^* completes μ so that together they form the maximal non-crossing partition of the doubled set. It is unique. Crossing partitions do not have a dual complement. The Kreweras complement defines an order-reversing bijection on $NC(n)$; that is, if $\nu \leq \mu$ then $\nu^* \geq \mu^*$. In other words, the coarsest partitions become the finest under the Kreweras complement, the finest become the coarsest, and intermediate levels are

mirrored. As a consequence, the Kreweras complement realizes an isomorphism between $NC(n)$ and the order-reversed lattice. μ^* is also called the dual partition of μ , and $NC(n)$ is said to be self-dual.

d. Free Cumulants. The combinatorics of partitions plays a central role in probability theory when defining cumulants recursively from the moments of a random variable A , through the moments-cumulants formula. In the classical case all partitions contribute, but in free probability theory only non-crossing partitions appear. The moments-free cumulants formula then reads

$$\langle A^n \rangle = \sum_{\mu \in NC(n)} \kappa_\mu, \quad \kappa_\mu = \prod_{m=1}^{\#\mu} \kappa_{l_m}, \quad (17)$$

where κ_μ is defined implicitly for each $\mu \in NC(n)$. Each κ_μ can be expressed as a product of free cumulants associated with the blocks of μ : l_m denotes the length of the m th block B_m of μ . As an example, the first free cumulants are defined implicitly by

$$\begin{aligned} \langle A \rangle &= \kappa_1 \\ \langle A^2 \rangle &= \kappa_2 + \kappa_1^2 \\ \langle A^3 \rangle &= \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3 \\ \langle A^4 \rangle &= \kappa_4 + 4\kappa_1\kappa_3 + 2\kappa_2^2 + 6\kappa_1^2\kappa_2 + \kappa_1^4. \end{aligned} \quad (18)$$

This formulation highlights the combinatorial structure underlying free cumulants: each cumulant naturally corresponds to a non-crossing partition of the indices, and the n th moment is reconstructed by summing over the contributions associated with all non-crossing partitions of size n . The moment functional $\langle \bullet \rangle$ can be chosen quite generally; in RMT, a common and natural choice is the normalized trace $\langle \bullet \rangle = \text{Tr}(\bullet)/D$.

e. Isomorphism partitions-permutations. This correspondence is discussed in more detail in Appendix A 3, while here we only outline the main idea. Every permutation in the symmetric group $\alpha \in S_n$ induces a partition μ_α by grouping together the elements belonging to the same cycle: $\mu_\alpha = \{\text{cycles of } \alpha\}$. The converse is not true: a partition does not specify a unique permutation, because one must choose an order of the elements within each block to define a cycle.

When fixing an order in the blocks, which for us will be the normal ordering, and excluding the crossing partitions, the correspondence becomes bijective. Therefore, one can establish an isomorphism between non-crossing partitions and a specific subset of permutations, the ones lying on the geodesic interval from the identity to the normal-ordered maximal cycle, where the geodesic is defined with respect to the Cayley distance. This statement is made more precise in Appendix A 3, where also the correspondence between the lattice structure and the Cayley graph is highlighted, after introducing a partial order on the permutations. Acknowledging this isomorphism, we will identify a non-crossing partition with the associated permutation. In particular, we will call them with the same letter, and represent blocks and cycles in the same way. The interpretation depends on the context: when viewed as a map, the notation indicates the action of a permutation, otherwise, it represents the blocks of a partition. We indicate as $e = (1) \cdots (n)$ the identity permutation i.e. the discrete partition, and with $\gamma_n = (1 \cdots n)$ the maximal normal-ordered cycle i.e. the trivial partition.

III. GLOBAL ROTATIONAL INVARIANCE

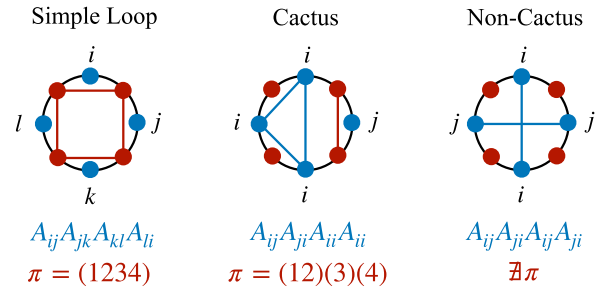
Our goal is to evaluate the products in Eq. (11) and derive a formula that provides a combinatorial prescription for computing the leading order contribution and the first subleading, in the Hilbert space dimension D . In this section, we consider the case in which the observable A^U possesses global rotational invariance. The results presented here are derived in Appendix B and are obtained by inserting Eq. (12) into Eq. (11) and performing a systematic expansion in powers of $1/D$. Our analysis focuses on specific products of the form $\overline{A_{i_1 i_2}^U} \overline{A_{i_2 i_3}^U} \cdots \overline{A_{i_n i_1}^U}$, which we refer to as *ETH-cycles* [84]. For convenience, we define the ordered set of indices $\mathbf{i} := \{i_1, \dots, i_n\}$, where different i_j may or may not coincide, for different j . By allowing all possible repetitions in $\mathbf{i}$, we encompass all the relevant products of matrix elements that we aim to study. Each can be uniquely associated with a partition of the set $\{1 \cdots n\}$, where the numbers indicate the position of indices in $\mathbf{i}$, meaning the subindices: we allocate in the same block the positions of equal indices in the set $\mathbf{i}$. Specifically, let B_m denote the m th block of the partition, containing l_m elements. All indices corresponding to that block are

equal, i.e. $i_{B_m(1)} = \cdots = i_{B_m(l_m)}$. The total number of distinct indices in the product is then given by the number of blocks in the partition. For instance, $A_{ij}A_{ji}A_{ii}A_{ii}$ has $\mathbf{i} = \{i, j, i, i\}$ and it is associated with the 2-blocks partition $(134)(2)$. Indeed, $i_1 = i_3 = i_4 \equiv i \neq i_2 \equiv j$, with respect to the generic ETH-cycle $A_{i_1 i_2} A_{i_2 i_3} A_{i_3 i_4} A_{i_4 i_1}$.

Based on which indices are repeated, we can have:

- (i) *Simple loops*, if they are all different. These are defined by the n -blocks partitions;
- (ii) *Cacti*, if there are repeated indices and the associated partition is non-crossing;
- (iii) *Non-cacti*, if there are repeated indices and the associated partition is crossing. For instance, this is the case for $A_{ij}A_{ji}A_{ij}A_{ji}$ with partition $(13)(24)$. We will exclude non-cacti diagrams from our discussion [85].

Each ETH-cycle is associated with a diagram that can be identified with the one of the partition so defined [32]. For instance, some examples for $n = 4$ are



In this diagrammatic representation, we highlight the original partition of the indices in blue and its dual non-crossing partition (when it exists) in red. In fact to write the closed formula for leading and subleading contributions, it is convenient to classify the diagrams in terms of the Kreweras complement of the partition introduced before, which we denote generically by π . While the partition introduced previously ensured that indices labelled by numbers in the same block were equal, for π this is not the case; it is more convenient to interpret it as a permutation. Concretely, π can be seen as a permutation of the set of subindices $\{1, \dots, n\}$ of $\mathbf{i}$; while $\mathbf{i} = \{i_1, \dots, i_n\}$ is the set of the indices in the first position in the matrix elements, $\pi(\mathbf{i}) := \{i_{\pi(1)}, \dots, i_{\pi(n)}\}$ is the set of indices in the second position in the matrix elements. This allows us to express a product of matrix elements, excluding non-cacti, in the form $A_{i_1 i_{\pi(1)}} \cdots A_{i_n i_{\pi(n)}}$. For example, $A_{ij}A_{ji}A_{ii}A_{ii}$ has Kreweras complement $\pi = (12)(3)(4)$ on $\mathbf{i} = \{i, j, i, i\}$, which reflects the fact that the first two indices form a cycle and are exchanged in the first two matrix elements: $i_{\pi(1)} = i_2 = j$ and $i_{\pi(2)} = i_1 = i$, while $i_{\pi(3)} = i_3 = i$ and $i_{\pi(4)} = i_4 = i$. For cacti, the number of blocks $\#\pi$ of the dual partition is the number of *leaves* in the diagram,

meaning the number of parts in which the circle is divided when *pinching* equal indices.

With the tools introduced so far, we can now explain the formula for leading and (first) subleading contributions of $\overline{A_{i_1 i_{\pi(1)}}^U \cdots A_{i_n i_{\pi(n)}}^U}$. This is given here as an ansatz, which is proven in Appendix B:

$$\overline{A_{i_1 i_{\pi(1)}}^U \cdots A_{i_n i_{\pi(n)}}^U} = \frac{\kappa_\pi}{D^{n-\#\pi}} + \sum_{\substack{\tau \in NC(n) \text{ with} \\ \tau > \pi, \#\tau = \#\pi - 1}} \frac{\kappa_\tau}{D^{n-\#\tau}} + \mathcal{O}\left(\frac{1}{D^{n-\#\pi+2}}\right), \quad (19)$$

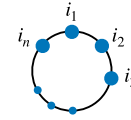
where $\kappa_\mu = \prod_{m=1}^{\#\mu} \kappa_{I_m}$ is a product of free cumulants, one for each block, being $\#\mu$ the number of blocks of μ , cf. Eq. (17). These free cumulants are the ones associated with moments given by the normalized traces of powers of the observable A . We emphasize that, in Eq. (19), π does not have to be thought as a generic permutation, but as the one associated with the Kreweras complement of the non-crossing partition describing simple loops and cacti.

Let us now discuss the implications of Eq. (19). The leading term contribution is given by the product of free cumulants associated with the blocks of π , rescaled by the Hilbert space dimension to a power dependent on the number of blocks. This constitutes the simplest “toy model” of ETH (see for instance [7] for $n = 2$), in which products of matrix elements have expectation values which do not depend on the indices, the constant free cumulants of Eq. (18). The leading term was obtained in a different context and with the methods of Appendix E in [86]. Remarkably, even the subleading correction admits a compact expression in terms of free cumulants, the ones associated with partitions τ bigger than π at distance one on the lattice of non-crossing partitions. Therefore, the subleading contributions can be read immediately from the lattice $NC(n)$, locating π , and considering all the partitions on the row above, connected with a line to it. The partitions τ may, depending on π , be unique, non-unique, or may not exist.

Let us illustrate these results by explicit examples. For simple loops, the partition τ at distance one does not exist. In fact, π is the trivial and the coarsest partition possible: it is located at the top of the lattice, and there are no partitions above. Hence, the first-order correction is zero [87]. Its associated free cumulant is $\kappa_\pi = \kappa_n$. For instance, for $\pi = (1234)$, partition at the top of $NC(4)$ in Fig. 3, the formula reads:

$$\overline{A_{ij}^U A_{jk}^U A_{kl}^U A_{li}^U} = \frac{\kappa_4}{D^3} + \mathcal{O}\left(\frac{1}{D^5}\right). \quad (20a)$$

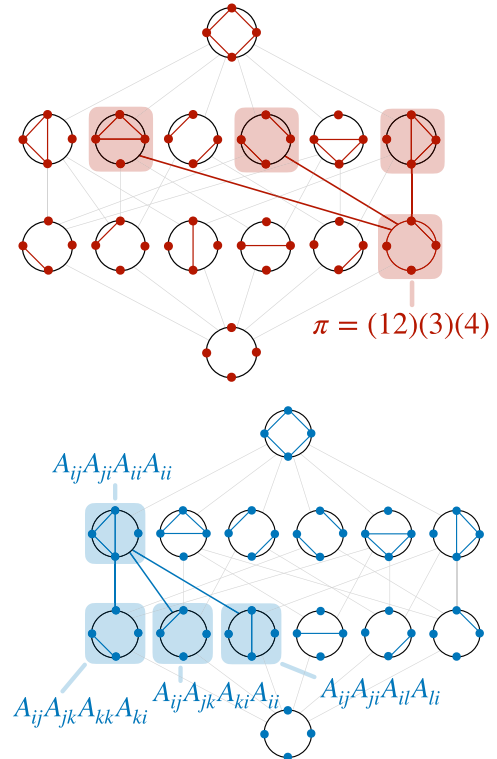
As a consequence, one can identify a free cumulant with a diagram with all distinct indices as

$$\kappa_n / D^{n-1} \simeq \overline{A_{i_1 i_2}^U A_{i_2 i_3}^U \cdots A_{i_n i_1}^U}. \quad (20b)$$


Throughout this work, the symbol $\simeq$ denotes the leading and next-to-leading behavior, while all corrections $\mathcal{O}(D^{-2})$ smaller by two orders of magnitude are neglected.

For the case of cacti, the formula gives the factorization of the matrix elements' n th moments in terms of lower order moments with different indices only, each of them given by a single free cumulant κ , and each of them representing a block of the initial diagram. This is the usual factorization encoded by full ETH: one should divide the statistical average every time a repeated index is present. The subleading contribution is a correction to the standard factorization. This can be understood more clearly in the following examples.

For $\pi = (12)(3)(4)$, Kreweras complement of $\overline{A_{ij}^U A_{ji}^U A_{ii}^U A_{ii}^U}$, the leading order is given by $\kappa_2 \kappa_1^2 / D$, which corresponds to the factorization $\overline{A_{ij}^U A_{ji}^U A_{ii}^U A_{ii}^U}$. For the subleading, there are three non-crossing partitions with $\#\tau = 2$, which are $(12)(34)$, $(123)(4)$ and $(124)(3)$. From the lattices,



We represented both the lattice for partition associated with the repeated indices in a product of matrix elements (second lattice) and the lattice for its Kreweras complement (first lattice). Indeed, since the lattice is self-dual, one

can read the result indifferently from one or the other, just reversing the order. From the lattice on the bottom, it appears more clearly that the partitions finer of one unity than π correspond to products of matrix elements in which there is one additional distinct index, compared with the initial product. Graphically, this corresponds to *breaking the lines* in the initial block one at a time, which effectively means considering one fewer repeated index and thus one fewer block. The subleading corrections are given by the leading contributions of these configurations. As a result:

$$\begin{aligned} \overline{A_{ij}^U A_{ji}^U A_{ii}^U A_{ii}^U} &= \frac{\kappa_2 \kappa_1^2}{D} + \frac{\kappa_2^2 + 2\kappa_1 \kappa_3}{D^2} + \mathcal{O}\left(\frac{1}{D^3}\right) \\ &\simeq \overline{A_{ij}^U A_{ji}^U} \overline{A_{ii}^U} \overline{A_{ii}^U} + \overline{A_{ij}^U A_{ji}^U} \overline{A_{il}^U A_{li}^U} + 2\overline{A_{ij}^U A_{jk}^U A_{ki}^U} \overline{A_{ii}^U} \\ &\simeq \text{diagrams} \end{aligned} \quad (20c)$$

The subleading is made by *recombining* the factorized terms of the leading order in all possible ways, while treating indices that were equal as distinct; these indices are connected by a dashed line in the diagrams.

Consider as another example $\pi = (12)(34)$, the Kreweras complement of $\overline{A_{ij}^U A_{ji}^U A_{il}^U A_{li}^U}$: the leading order is given by κ_2^2/D^2 , which corresponds to the factorization $\overline{A_{ij}^U A_{ji}^U} \overline{A_{il}^U A_{li}^U}$. For the subleading, there is just one non-crossing partition with $\#\tau = 1$, which is γ_4 , the one at the top of the dual lattice. This corresponds, indeed, to the product of matrix elements in which there is one additional distinct index, compared with the initial product. As a result:

$$\begin{aligned} \overline{A_{ij}^U A_{ji}^U A_{il}^U A_{li}^U} &= \frac{\kappa_2^2}{D^2} + \frac{\kappa_4}{D^3} \\ &\simeq \overline{A_{ij}^U A_{ji}^U} \overline{A_{il}^U A_{li}^U} + \overline{A_{ij}^U A_{jk}^U A_{kl}^U A_{li}^U} \\ &\simeq \text{diagrams} \end{aligned} \quad (20d)$$

Therefore, our finding is that *each subleading contribution of cacti equals the leading contribution of a configuration with one more distinct index, previously coincident in the cacti*. For non-cacti, the partition π

is not defined which reflects the fact that there is no unique way of factorizing these products of matrix elements, as it is better understood in the derivation in Appendix B.

The expression in Eq. (19) contains no explicit energy dependence; indeed, it is well known that this simple RMT toy model is not sufficient to derive the ETH functions. In the following section, we extend this approach in order to incorporate the explicit energy dependence and overcome this limitation.

IV. LOCAL ROTATIONAL INVARIANCE

We now consider the second toy model, which takes into account that A remains invariant under local basis transformations, i.e., reshufflings of its elements restricted to local neighborhoods, i.e. $A \sim A^{U_{\text{loc}}}$. We consider U_{loc} to have local structure in the energy of the Hamiltonian, as explained in Sec. II. The Haar average for computing the matrix elements' n th moments (16) is performed explicitly in Appendix C, for some examples involving the first few n orders. Although the derivations are made using the minimal model in Fig. 2(b), the resulting expressions can be directly translated to the more refined model in Fig. 2(c). Here we present these results, separating the discussion between simple loops and cacti.

A. Local free cumulants

For simple loops, we show in Appendix C that the average is given by a smooth function of the energies,

$$\overline{A_{i_1 i_2}^U \cdots A_{i_n i_1}^U} \simeq K_n(E_{i_1}, \dots, E_{i_n}), \quad (21)$$

that we now write explicitly for the first few orders up to $n = 4$. To specify the functions K_n , let us introduce a normalized delta operator, for each energy state i ,

$$\hat{\sigma}_{E_i} \equiv \sum_{\alpha} \frac{\delta_{\Delta}(E_i - E_{\alpha})}{D_i} |\alpha\rangle\langle\alpha|, \quad \text{Tr}(\hat{\sigma}_{E_i}) = 1, \quad (22)$$

where we recall that $\delta_{\Delta}(E_i - E_{\alpha})$ is peaked around E_i with a characteristic width Δ , so that $\hat{\sigma}_{E_i}$ implements a smooth microcanonical projector on energy E_i . We denote with $\sigma(E_i - E_{\alpha}) = \delta_{\Delta}(E_i - E_{\alpha})/D_i$ the diagonal matrix element of such operator. This allows to write the smooth functions K_n as

$$K_1(E_i) = \text{Tr}(\hat{\sigma}_{E_i} A) \quad (23a)$$

$$K_2(E_i, E_j) = \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_j} A) - [\text{Tr}(\hat{\sigma}_{E_i} A)]^2 \sigma(E_i - E_j) \quad (23b)$$

$$K_3(E_i, E_j, E_k) = \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_j} A \hat{\sigma}_{E_k} A) - \text{Tr}(\hat{\sigma}_{E_i} A) \sigma(E_i - E_j) \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_k} A) + 2 \text{ cycl.} \\ + 2 \text{Tr}(\hat{\sigma}_{E_i} A)^3 \sigma(E_i - E_j) \sigma(E_j - E_k) \quad (23c)$$

$$K_4(E_i, E_j, E_k, E_l) = \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_j} A \hat{\sigma}_{E_k} A \hat{\sigma}_{E_l} A) \\ - \text{Tr}(\hat{\sigma}_{E_i} A) \sigma(E_i - E_j) \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_k} A \sigma_{E_l} A) + 3 \text{ cycl.} \\ - \text{Tr}(\hat{\sigma}_{E_i} A \sigma_{E_j} A) \sigma(E_i - E_k) \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_l} A) + \text{cycl.} \\ + 2 \text{Tr}(\hat{\sigma}_{E_i} A)^2 \sigma(E_i - E_j) \sigma(E_j - E_k) \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_l} A) + 3 \text{ cycl.} \\ + \text{Tr}(\hat{\sigma}_{E_i} A) \text{Tr}(\hat{\sigma}_{E_k} A) \sigma(E_i - E_j) \sigma(E_k - E_l) \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_k} A) + \text{cycl.} \\ - 5 \text{Tr}(\hat{\sigma}_{E_i} A)^4 \sigma(E_i - E_j) \sigma(E_j - E_k) \sigma(E_k - E_l). \quad (23d)$$

It is straightforward to see that for a flat $\delta_\Delta() = 1$ over the entire energy spectrum, one recovers the results of global rotational invariance (20b), discussed in Sec. III. In this limit, these correlators correspond to rescaled free cumulants, $K_n(E_{i_1}, \dots, E_{i_n}) = \kappa_n / D^{n-1}$, where κ_n were defined from normalized traces of powers of the entire matrix, cf. Eq. (18).

We assume that, when the smooth functions are multiplied by the energy projectors, then they can be evaluated at the same energy, i.e. $K_1(E_i) \sigma(E_i - E_j) = K_1(E_j) \sigma(E_i - E_j)$. This implies the continuity of the smooth functions when the indices are the same (zero frequencies).

By inverting the (23) we find:

$$\text{Tr}(\hat{\sigma}_{E_i} A) = K_1(E_i) \quad (24a)$$

$$\text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_j} A) = K_2(E_i, E_j) + K_1(E_i) K_1(E_j) \sigma(E_i - E_j) \quad (24b)$$

$$\text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_j} A \hat{\sigma}_{E_k} A) = K_3(E_i, E_j, E_k) + K_1(E_i) K_2(E_j, E_k) \sigma(E_i - E_j) + 2 \text{cycl} \\ + K_1(E_i) K_1(E_j) K_1(E_k) \sigma(E_i - E_j) \sigma(E_j - E_k) \quad (24c)$$

$$\text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_j} A \hat{\sigma}_{E_k} A \hat{\sigma}_{E_l} A) = K_4(E_i, E_j, E_k, E_l) + K_1(E_i) K_3(E_j, E_k, E_l) \sigma(E_i - E_j) + 3 \text{ cycl} \\ + K_2(E_i, E_j) K_2(E_k, E_l) \sigma(E_i - E_k) + \text{cycl} \\ + K_1(E_i) K_1(E_j) K_2(E_k, E_l) \sigma(E_i - E_j) \sigma(E_j - E_k) + 3 \text{ cycl} \\ + K_1(E_i) K_1(E_k) K_2(E_j, E_l) \sigma(E_i - E_j) \sigma(E_k - E_l) + \text{cycl} \\ + K_1(E_i) K_1(E_j) K_1(E_k) K_1(E_l) \sigma(E_i - E_j) \sigma(E_j - E_k) \sigma(E_k - E_l). \quad (24d)$$

One can notice the similarity between these expressions and Eq. (18), in fact, K_n can be defined implicitly from a generalized moments-cumulants formula as

$$\text{Tr}(\hat{\sigma}_{E_{i_1}} A \cdots \hat{\sigma}_{E_{i_n}} A) \\ = \sum_{\mu \in NC(n)} K_\mu(E_{i_1}, \dots, E_{i_n}) \sigma_{\mu^*}(E_{i_1}, \dots, E_{i_n}), \quad (25a)$$

where μ^* is the dual of the non-crossing partition μ , and the function $K_\mu(E_{i_1}, \dots, E_{i_n})$ is defined as

$$K_\mu(E_{i_1}, \dots, E_{i_n}) = \prod_{m=1}^{\#\mu} K_{l_m}(E_{i_{B_m(1)}}, \dots, E_{i_{B_m(l_m)}}). \quad (25b)$$

The function $\sigma_{\mu^*}(E_{i_1}, \dots, E_{i_n})$ is a product of rescaled delta functions $\sigma()$ that equate indices belonging to the

same block of μ^* ; note that if multiple indices belong to the same block we should choose the minimum number of $\sigma()$ functions. By inspection of the first few orders, one immediately finds Eq. (24).

As it is clear from Eq. (23), when the energies are well separated, these quantities can be rewritten as traces with appropriate normalization, so that their scaling reads:

$$K_n(E_1, \dots, E_n) \simeq D_{E_+}^{1-n}, \quad (26)$$

where on the right-hand side, we take the large-size limit $L \gg 1$ and use that for a finite-rank observable, the trace scales as the density of states at the average energy $E_+ = (E_1 + \dots + E_n)/n$.

We note that, besides dimensional factors, the same expression (25) can be found in Refs. [60,73,88], where

the corresponding quantities are defined as *local free cumulants*. Therefore, in this work we will likewise refer to K_n as local free cumulants. The main difference is that our K_n , derived within the local rotational invariant model, possess the different dimensionality associated with the energies $D_i, D_j, \dots$. In Ref. [60] the density of states of the matrices is flat: there is no energy dependency and all the D are the same, as for Floquet models, so one can define properly normalized functions. Moreover, Ref. [60] shows that the local free cumulants introduced there can be related to operator-valued free cumulants.

Since Eq. (25) are the generalization of Eq. (17) to matrices with an energy structure, this holds also in our case: we show in Appendix D how Eq. (25) naturally appear when generalizing scalar free probability to operator-valued free probability. Consequently, local free cumulants can be identified with the diagonal components of specific operator-valued free cumulants, namely, those weighted by projectors onto energy windows.

B. Cacti diagrams

Finally, let us discuss the implications for the scaling of matrix elements with repeated indices. For these diagrams, one can characterize not only the leading-order, but also the subleading scaling, as we did for the model with global rotational invariance.

We illustrate here examples; for $n = 2$, the expectation value with repeated indices reads

$$\overline{A_{ii}^U A_{ii}^U} \simeq [K_1(E_i)]^2 + K_2(E_i, E_i) \simeq \overline{A_{ii}^U}^2(E_i) + \overline{A_{ij}^U A_{ji}^U}(E_i, E_i). \quad (27a)$$

This corresponds to the usual factorization with correction given by the second local free cumulant evaluated at equal energies. In Eq. (27) and in the following, we use the notation $\overline{A_{ij}^U A_{ji}^U}(E_i, E_i)$ to highlight the energy dependencies. This result for the diagonal matrix elements matches the well known ETH two-point function result [89].

For $n = 4$, two examples are:

$$\begin{aligned} \overline{A_{ij}^U A_{ji}^U A_{ii}^U A_{ii}^U} &\simeq K_2(E_i, E_j)[K_1(E_i)]^2 \\ &+ K_2(E_i, E_j)K_2(E_i, E_i) + 2K_1(E_i)K_3(E_i, E_j, E_i) \\ &\simeq \overline{A_{ij}^U A_{ji}^U}(E_i, E_j)\overline{A_{ii}^U}^2(E_i) + \overline{A_{ij}^U A_{ji}^U}(E_i, E_j)\overline{A_{ii}^U A_{ii}^U}(E_i, E_i) \\ &+ 2\overline{A_{ii}^U}(E_i)\overline{A_{ij}^U A_{jk}^U A_{kl}^U}(E_i, E_j, E_i), \end{aligned} \quad (27b)$$

and

$$\begin{aligned} \overline{A_{ij}^U A_{ji}^U A_{il}^U A_{li}^U} &\simeq K_2(E_i, E_j)K_2(E_i, E_l) + K_4(E_i, E_j, E_i, E_l) \\ &\simeq \overline{A_{ij}^U A_{ji}^U}(E_i, E_j)\overline{A_{il}^U A_{li}^U}(E_i, E_l) \\ &+ \overline{A_{ij}^U A_{jk}^U A_{kl}^U A_{li}^U}(E_i, E_j, E_i, E_l), \end{aligned} \quad (27c)$$

from which it is clear that, at the leading order, the usual factorization happens, while the corrections arise from higher-order local free cumulants, corresponding to products of matrix elements with more different indices, but evaluated at equal mesoscopic energies.

These results generalize those obtained for global rotational invariance in Sec. III to the case of structured matrices. In particular, Eqs. (27b)–(27c) are analogous to Eqs. (20c)–(20d), but now incorporating explicit energy dependence. One could notice that the dashed lines in the diagrams above are, in the case of local rotational invariance, connecting equal mesoscopic energies.

Therefore, the contributions can be read again from the lattice of non-crossing partitions, and the general expression for the leading and subleading behavior reads:

$$\overline{A_{i_1 i_{\pi(1)}}^U \cdots A_{i_n i_{\pi(n)}}^U} \simeq K_\pi(E_{i_1}, \dots, E_{i_n}) + \sum_{\substack{\tau \in NC(n) \text{ with} \\ \tau > \pi, \# \tau = \# \pi - 1}} K_\tau(E_{i_1}, \dots, E_{i_n}), \quad (28)$$

where, in the set $\{E_{i_1}, \dots, E_{i_n}\}$ the energies are repeated whenever indices are repeated in $\mathbf{i}$; the function K_μ was defined in Eq. (25b), as a product of local free cumulants. For K_π , even if in $\{E_{i_1}, \dots, E_{i_n}\}$ there can be repetitions, in the dependencies of its local free cumulants $K_{l_m}(\pi)$ only distinct energies are present, since, by definition, each block of π accounts only for different indices. For K_τ , instead, since in the products associated there is one additional distinct index compared with the initial product, $K_{l_m}(\tau)$ is calculated at equal energies whenever indices that were previously identical become distinct.

As a result, *each subleading contribution of cacti equals the leading contribution of a configuration with one more distinct index, calculated at equal energies for the indices previously coincident.*

Equation (28) represents one of the main results of this paper, and it directly generalizes the result of full RMT in Eq. (19) to structured matrices. The same behavior is expected to hold for matrix elements of non-integrable Hamiltonians, hence providing a refinement of the full ETH. In Sec. VI, we shall verify that the same also holds for a non-integrable Floquet model.

V. EMPIRICAL AVERAGES OVER SMALL ENERGY WINDOWS

After the local rotational invariant toy model, we are now in position to make a connection with the standard ETH ansatz. Firstly, the scaling in Eq. (26) of the local free cumulants K_n suggests defining some order-one functions

$$F_{e_+}^{(n)}(\omega_1, \dots, \omega_{n-1}) = \lim_{L \rightarrow \infty} e^{(n-1)Ls(e_+)} K_n(E_1, \dots, E_n), \quad (29)$$

commonly known as *ETH smooth functions*, which are generally written as functions of the mean energy density $e_+ = E_+/N$ and of all the other $n-1$ independent frequencies $\omega_i = E_i - E_{i+1}$.

In order to make a connection with the ETH ansatz in Eq. (1), an important step forward that we make in this work is to link the previously defined average of matrix elements over an ensemble of banded unitaries to the *empirical averages* of matrix elements of a given instance of the operator under study, in the basis of the particular Hamiltonian that one has chosen. We can derive, for diagrams with distinct indices:

$$\begin{aligned} K_n(E_{i_1}, \dots, E_{i_n}) &= \overline{A_{i_1 i_2}^U \dots A_{i_n i_1}^U} \\ &\simeq \sum_{\alpha_1 \neq \dots \neq \alpha_n} A_{\alpha_1 \alpha_2} \frac{\delta_\Delta(E_{i_1} - E_{\alpha_1})}{D_{i_1}} \dots A_{\alpha_n \alpha_1} \frac{\delta_\Delta(E_{i_n} - E_{\alpha_n})}{D_{i_n}} \\ &\equiv \overline{A_{i_1 i_2} \dots A_{i_n i_1}} \quad \text{for } i_1 \neq \dots \neq i_n. \end{aligned} \quad (30)$$

In order to derive this relation we have assumed that the statistical properties of A are the same as those of A^U so that in the *generalized moments* in (23) or in the generalized sums in (30) we could arbitrarily use A or A^U . We have then assumed that these quantities are self-averaging, which is typically the case in RMT. This expression corresponds just to a definition of average over smoothed different indices for a single realization of the matrix; this is what we define as empirical average. The nice aspect of this equation is that it corresponds to the numerical recipe for deterministic matrix elements, as commonly used in the literature [33]. Furthermore, this expression is continuous for coinciding energies, i.e. when $E_i \rightarrow E_j$ or $\omega_{ij} = E_i - E_j \rightarrow 0$.

Similarly, one can define empirical averages of diagrams with repeated indices. The strategy is to average the matrix elements over small windows of energies, keeping the sum and the smoothing only for the distinct indices. As an example we consider:

$$\overline{A_{ii} A_{ii}}(E_i) \equiv \sum_{\alpha} A_{\alpha\alpha} A_{\alpha\alpha} \frac{\delta_\Delta(E_i - E_\alpha)}{D_i}. \quad (31)$$

Under the assumptions made above one can recognize these empirical averages as the averages over the ensemble of unitaries and obtain the factorization and its corrections. Other examples of empirical averages can be found in (F2).

VI. NUMERICAL TEST IN FLOQUET MODELS

This section presents a numerical test of both the well established ETH and the new subleading corrections derived analytically in the previous sections. We consider a non-integrable one-dimensional spin chain under a stroboscopic periodic drive, a setup similar to those commonly studied in the literature on Floquet many-body systems [90,91] and in recent works on quantum chaos

[33,92–94]. We introduce the model below and subsequently present the simulation methods and results. All numerical codes and datasets used in this work are openly available on Zenodo [95].

A. The model

We consider the following time-dependent Hamiltonian

$$H(t) = H_0 + \text{sgn}[\cos(\Omega t)] V_0, \quad (32)$$

with H_0 and V_0 accounting for (i) Ising interactions on a chain of L spin 1/2, (ii) longitudinal and transverse fields, and (iii) mixed interactions between nearest neighbors in the transverse directions, similar to the Dzyaloshinskii-Moriya interactions. Specifically:

$$\begin{aligned} H_0 &= H_Z + H_X + H_M, & V_0 &= H_Z - H_X - H_M, \\ H_Z &= J \sum_i Z_i Z_{i+1} + h \sum_i Z_i, & H_X &= g \sum_i X_i, \\ H_M &= d \sum_i (X_i Y_{i+1} + Y_i X_{i+1}). \end{aligned} \quad (33)$$

We consider periodic boundary conditions. The time evolution operator for (32), over one period T , is

$$\begin{aligned} U_F &= e^{-i(H_0+V_0)\frac{T}{4}} e^{-i(H_0-V_0)\frac{T}{2}} e^{-i(H_0+V_0)\frac{T}{4}} \\ &= e^{-iH_Z \frac{T}{2}} e^{-i(H_X+H_M)T} e^{-iH_Z \frac{T}{2}} \\ &\equiv \sum_{\alpha} e^{-i\nu_{\alpha}} |\nu_{\alpha}\rangle \langle \nu_{\alpha}|, \end{aligned} \quad (34)$$

whose spectrum is given by the Floquet quasienergies ν_{α} , which are 2π periodic; the choice $\nu_{\alpha} \in [-\pi, \pi)$, $\forall \alpha$ is made. The frequency of the drive Ω is connected to the period as $T = 2\pi/\Omega$. For definiteness, we fix $\Omega = 2\pi$, and the parameters $J = 1$, $h = 0.905$, $g = 0.809$, $d = 1$.

This ergodic model exhibits translational invariance and parity symmetry, but not time-reversal symmetry, which is broken by the H_M interaction [96]. It therefore belongs to the unitary symmetry class, meaning that the level-spacing statistics of the Floquet operator eigenvalues follow those of a random Hamiltonian drawn from the Gaussian unitary ensemble (GUE) [98]. This is precisely the case for which we derived the analytical results: the observable A was required to be invariant under unitary rotations, reflecting that the Floquet eigenvectors belong to the unitary class. As observable, we choose the total magnetization, rescaled by the number of spins, in order to preserve translational invariance.

In Appendix F we present results for slightly different Hamiltonians. Firstly, we consider a Hamiltonian without translational invariance, in order to check our assumptions for different observables, for instance, the magnetization on a single site. Then, we consider the case for which the

time reversal symmetry is restored, and the Hamiltonian belongs to the orthogonal symmetry class; we will briefly discuss the discrepancies observed with respect to the unitary case.

B. Methods

Our aim is to prove numerical Eq. (28), in particular we will look at two specific matrix elements' 4th moments that are the ones in the examples (27b)–(27c). We want to verify the validity of the factorization and its subleading corrections, when calculating numerically the product of matrix elements through the empirical averages. Moreover, we want to confirm the scaling with the Hilbert (sub)space dimension of the different contributions.

We begin by specializing our findings to Floquet models. In this case, the ETH smooth functions do not depend on the absolute energies, but only on the energy differences [99]. $\omega_{\alpha\beta} = \nu_\alpha - \nu_\beta$. This reflects the fact that in periodically driven systems there is no conserved total energy, and thus the statistical properties of matrix elements can only depend on relative quasienergies separations. Moreover, the density of states is flat so that $D_{E^+} = e^{Ls(e^+)} = D$, Hilbert (sub)space dimension. Therefore, for simple loops, calling $\omega_\alpha \equiv \omega_{i_\alpha i_{\alpha+1}}$,

$$\overline{A_{i_1 i_2} \cdots A_{i_n i_1}}(\omega_1, \dots, \omega_{n-1}) \simeq F^{(n)}(\omega_1, \dots, \omega_{n-1})/D^{n-1}. \quad (35)$$

The relation is similar to the one of global rotational invariance, while still keeps the dependence on the difference of quasienergies.

Let us write the expressions for the cacti in (27b)–(27c), specialized for the Floquet case.

$$\begin{aligned} \overline{A_{ij} A_{ji} A_{ii} A_{ii}}(\omega) &\simeq \overline{A_{ij} A_{ji}}(\omega) \overline{A_{ii}^2} \\ &+ \overline{A_{ij} A_{ji}}(\omega) \overline{A_{il} A_{li}}(0) + 2 \overline{A_{ii}} \overline{A_{ij} A_{jk} A_{ki}}(\omega, -\omega), \end{aligned} \quad (36a)$$

with $\omega \equiv \omega_{ij} = \nu_i - \nu_j$, and

$$\begin{aligned} \overline{A_{ij} A_{ji} A_{il} A_{li}}(\omega_1, \omega_2) &\simeq \overline{A_{ij} A_{ji}}(\omega_1) \overline{A_{il} A_{li}}(\omega_2) \\ &+ \overline{A_{ij} A_{jk} A_{kl} A_{li}}(\omega_1, -\omega_1, \omega_2), \end{aligned} \quad (36b)$$

with $\omega_1 := \omega_{ij} = \nu_i - \nu_j$, $\omega_2 := \omega_{il} = \nu_i - \nu_l$.

The numerical calculation of each product of matrix elements is made by employing empirical averages, where the summing and the smoothing are performed directly on the quasienergy differences and involve only distinct indices. It is important to account for the intrinsic periodicity of the quasienergies, ensuring that the smoothing procedure is formulated in a correspondingly periodic manner. Specifically, since ν_α is identified with $\nu_\alpha \pm 2\pi$, we employ:

$$\begin{aligned} \delta_\omega^{\text{per}}(\nu_\alpha, \nu_\beta) &= \delta_\Delta(\omega - (\nu_\alpha - \nu_\beta)) \\ &+ \delta_\Delta(\omega + 2\pi - (\nu_\alpha - \nu_\beta)) + \delta_\Delta(\omega - 2\pi - (\nu_\alpha - \nu_\beta)). \end{aligned} \quad (37)$$

Therefore, for simple loops the empirical average reads:

$$\begin{aligned} &\overline{A_{i_1 i_2} \cdots A_{i_n i_1}}(\omega_1, \dots, \omega_{n-1}) \\ &\simeq \frac{1}{D} \sum_{\alpha_1 \neq \dots \neq \alpha_n} A_{\alpha_1 \alpha_2} \frac{\delta_{\omega_1}^{\text{per}}(\nu_{\alpha_1}, \nu_{\alpha_2})}{D_{\omega_1}} \cdots \\ &\cdots A_{\alpha_{n-1} \alpha_n} \frac{\delta_{\omega_{n-1}}^{\text{per}}(\nu_{\alpha_{n-1}}, \nu_{\alpha_n})}{D_{\omega_{n-1}}} A_{\alpha_n \alpha_1}, \end{aligned} \quad (38)$$

with $D_\omega = \sum_{\alpha, \beta} \delta_\omega^{\text{per}}(\nu_\alpha, \nu_\beta)/D$. Consequently the formulation follows for the cacti. All the expressions numerically implemented are explicitly reported in Appendix F 3.

C. Numerical results

We perform exact diagonalization of the Floquet operator (34), using the QuSpin package [100,101] for writing the different terms in (33). In particular, we restrict the Hilbert space to the symmetry block of positive parity and zero momentum; we indicate with D the dimension of this subspace. As observable, we choose the total magnetization along the z direction, rescaled by the number of spins, meaning $A = \sum_i Z_i/L$. After evaluating the matrix elements in the Floquet basis $A_{\alpha\beta} = \langle \nu_\alpha | A | \nu_\beta \rangle$, we calculate the averages of the products of matrix elements involved in Eqs. (36a)–(36b), through the expressions (F2), fixing, in the second example, $\omega_2 = \omega_1 \equiv \omega$. We choose 51 values of ω in $[-\pi, \pi]$. Numerically, we implement the smoothing through a Gaussian function

$$\delta_\Delta(\omega - (\nu_\alpha - \nu_\beta)) = \frac{1}{\sqrt{\pi}\Delta} \exp\left(-\frac{(\omega - (\nu_\alpha - \nu_\beta))^2}{\Delta}\right), \quad (39)$$

choosing Δ so that a sufficient number of states is satisfying the condition, still much smaller than D . Moreover, we average each matrix elements' moment over multiple values of Δ s, in order to have a final function with the respective error bars, for each value of ω . In Appendix F 3 we explicitly write the intervals chosen for the parameter Δ .

All results are plotted in Fig. 4. For a fixed number of spin $L = 16$, we compare the right-hand side (purple curve) of Eqs. (36a)–(36b) with the respective leading order (pink curve), and then we subtract the leading to the right-hand side, in order to compare this difference (brown curve) with the prediction for the subleading (orange curve). All curves are plotted as a function of ω . Moreover, in panels (c), (f), we study the scaling with the subspace dimension D for each of these curves, for a fixed $\omega^* \approx -2.13$. This value was

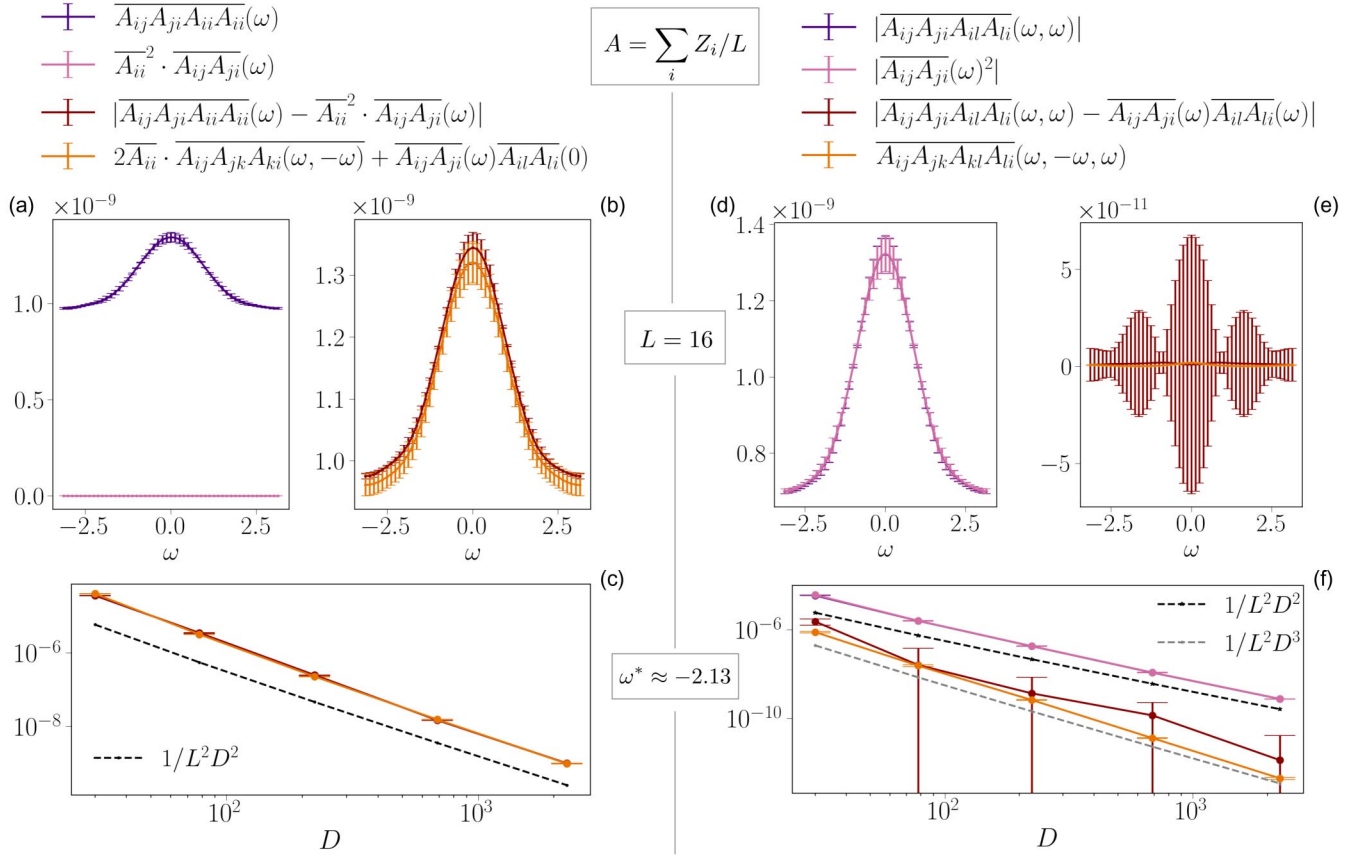


FIG. 4. Numerical results for the Floquet system in (34), and for the observable $A = \sum_i Z_i / L$. We numerically reproduce Eqs. (36a)–(36b), for $\omega_1 = \omega_2 = \omega$ chosen in $[-\pi, \pi]$. The error bars are given by averaging each product, for each ω , over different values of the smoothing parameter Δ , as explained in the text. (a), (b), (c) refer to (36a), while (d), (e), (f) to (36b). In (a), (d) we compare the right-hand side (purple curve) with the respective leading order factorization (pink curve) for fixed $L = 16$; for (a) the leading order is zero, while in (d) the data nicely reproduce the expected leading order factorization. In (b), (e) we compare the difference between the right-hand side and the leading contribution (brown curve) with the respective subleading correction (orange curve) for fixed $L = 16$. The data verify the subleading correction to the factorization. In (c), (f) we represent the scalings of all the curves as a function of D . The scalings were studied for values $L \in [8, 10, 12, 14, 16]$, and for a fixed value of $\omega^* \approx -2.13$. While all the curves in (c) scale as $1/L^2 D^2$, since the leading order is zero, in (f) we observe the different scalings between the leading (purple and pink curves), as $1/L^2 D^2$ and the subleading (brown and orange curves), as $1/L^2 D^3$.

chosen arbitrarily as a representative parameter value; the qualitative behavior of the plots is not sensitive to variations of ω^* which confirms the robustness of the results. The numerical results nicely agree the predictions, which follow from the fact that, for distinct indices, simple loops scale as $\overline{A_{i_1 i_2} \cdots A_{i_n i_1}} \propto 1/D^{(n-1)}$.

Since $\langle M_z \rangle = 0$, we notice that the leading order of $\overline{A_{ij} A_{ji} A_{ii} A_{ii}(\omega)}$ is zero, therefore the average follows directly the subleading contribution $\overline{A_{ij} A_{ji}(\omega) A_{ii} A_{ii}(0)}$, and scales as $1/L^2 D^2$, as expected. On the contrary, $\overline{A_{ij} A_{ji} A_{ii} A_{ii}(\omega)}$ is very well factorized in $\overline{A_{ij} A_{ji}(\omega) A_{ii} A_{ii}(\omega)}$, scaling as $1/L^2 D^2$, and the subleading well follows the expected curve $\overline{A_{ij} A_{jk} A_{kl} A_{li}(\omega, -\omega, \omega)}$, scaling as $1/L^2 D^3$. The rescaling with L is required since the magnetization has been rescaled with the total number of spins L , for being a local observable.

VII. CONCLUSIONS AND PERSPECTIVES

In this work we have revisited the general ideas behind the development of full ETH. In particular, we have discussed how the properties of matrix elements can be studied under random basis transformations which are associated with local perturbations of the Hamiltonian. Different levels of refinement of the model led us to consider (i) a completely random basis defined by the application of a Haar distributed random unitary matrix, (ii) a block diagonal unitary where each block is independent and Haar distributed, (iii) a smoothed version of this last model which can be associated with a banded diagonal unitary matrix.

The natural mathematical framework of our study is provided by free probability theory. In fact, for all these ensembles, we derived analytical expressions of moments of the matrix elements of the observable in terms of (local)

free cumulants. As it was noticed before, the products of interest, for the purposes of computing correlation functions, can be classified as diagrams such as “simple loops,” when indices concatenate but do not repeat, and “cacti,” which are non-crossing partitions of the indices with some repetitions. It was postulated that the expectation of cacti factorizes and in our work we provide a detailed characterization of the next to leading term beyond this assumption, which can be expressed in terms of non-crossing partitions at higher order.

In addition, we show how, under the assumption of (local) rotational invariance, the averages of the matrix elements under the application of random unitaries are compatible with the empirical averages performed over a single instance of the matrix ensemble, upon averaging over energy windows, the method that is usually performed in numerical simulations to extract the statistics of such entries.

As emphasized in the Introduction, deriving ETH and its subleading corrections from the principle of local rotational invariance provides a deeper understanding of the physical meaning of matrix-element correlations, of the averages entering the ETH ansatz, and of the ETH functions themselves in terms of global properties of the operator. At the same time, the characterization of

subleading corrections, already present in standard ETH, fills a structural gap in the full ETH framework.

More generally, these results open several promising directions for future research, which we discuss in the remainder of this section.

A. Applications of our results

We first discuss applications of the subleading corrections we have computed, which play a quantitative role in quantum dynamics.

(i) *Higher-order correlation functions: late-time plateaus and freeness.* The subleading corrections discussed in this paper determine the late-time plateau of *thermal free cumulants* [32]. In Appendix G we compute higher-order thermal correlation functions and explicitly identify the corresponding subleading terms at long times; these can be incorporated into thermal free cumulants, yielding a refinement of the ETH free cumulants introduced in Ref. [32]. Since the leading ETH free cumulants average to zero at late times due to the non-resonance condition [43], these corrections control the late-time plateau and the fluctuations around it. For the second and fourth thermal free cumulants, we find

$$[k_2(t, 0)]_\infty \simeq K_2(0), \quad (40)$$

$$[k_4(t, t, 0, 0)]_\infty \simeq \frac{1}{D} \sum_{i \neq j \neq l} K_4(\omega_{ij}, -\omega_{ij}, \omega_{il}) + \frac{1}{D} \sum_{i \neq j} K_2(\omega_{ij}) K_2(-\omega_{ij}), \quad (41)$$

of order $\mathcal{O}(D^{-1})$, with plateau fluctuations of order $\mathcal{O}(D^{-2})$. The subleading corrections thus set the exponentially small late-time plateau, providing a quantitative characterization of long-time freeness in chaotic dynamics.

We note that for non-Floquet systems, one should consider further corrections arising from the dependence of observables on the average energy, see e.g. Ref. [96].

(ii) *Breakdown of ETH.* Our subleading corrections may also offer a sharp diagnostic of chaos. At $k = 2$ they predict that the variance of the diagonal matrix elements equals the off-diagonal two-point function at zero frequency [Eq. (5)], decaying as $1/D$. In integrable systems, instead, the diagonal matrix elements fluctuate much more strongly, with a variance that decays only *polynomially* in the system size, as $1/L^a$ [103], a violation of this $k = 2$ prediction. The all-order corrections derived here could thus provide a precise way to distinguish integrable from non-integrable systems. Relatedly, recent works on the *breaking of ergodicity*, e.g. the “fading ergodicity” near the many-body-localized regime [104], address the anomalous

low-frequency scaling of the ETH functions, where our framework could find natural applications.

(iii) *Different symmetry classes.* The subleading corrections also depend on the symmetry class. Already at $k = 2$, the diagonal variance is fixed by the off-diagonal two-point function at zero frequency up to a class-dependent factor, one for the unitary class and two for the orthogonal one [7,48,89] [Eqs. (E10)–(F1)], as we verify numerically in Appendix F 2. The difference stems from rotational invariance: in the orthogonal class the observable is invariant under local *orthogonal* transformations, so the permutation Weingarten calculus used here no longer applies. A full prediction of the higher- k discrepancies between symmetry classes remains an open problem.

B. New directions

Beyond these applications, the principle of local rotational invariance suggests several directions where it could be applied more broadly.

(i) *Characterization of an ETH ensemble.* Much recent attention has been devoted to characterizing ensembles of unitaries $\mathcal{E}$ and their ability to reproduce the Haar ensemble. A major open challenge is to identify the ensemble appropriate to chaotic Hamiltonian time evolution,

$$\mathcal{E}_H = \{e^{-iHt}, t \sim \mathcal{P}\}, \quad (42)$$

where $\mathcal{P}$ is a suitable distribution of times. Since such an evolution conserves energy, and, in fact, every eigenstate, the Haar ensemble must be endowed with additional structure in order to capture the randomness of chaotic dynamics. Most efforts so far have focused either on ensembles of states, in the context of deep thermalization and Hilbert-space ergodicity [105–108], or on random Hamiltonians [109]. Local rotational invariance, through the associated block unitary U_{loc} , provides a first constructive definition of such an *ETH ensemble*: an ensemble of unitaries that, applied to the matrix elements of an observable, reproduces by construction the correlations dictated by full ETH, including both the smooth energy dependence and the subleading corrections. Making the relation between U_{loc} and the ensembles above precise remains an important open problem.

(ii) *Connection to eigenstate correlations.* Here we have applied local rotational invariance to the matrix elements of *observables*. There is, however, great interest in the correlations between the *eigenstates* themselves [22,28], which control information-theoretic quantities and the spreading of operators. Since such quantities can be written as traces of suitable projector operators, local rotational invariance could in principle be applied to those projectors as well, providing a route to derive eigenstate-correlation results within the same framework.

(iii) *Block and banded unitary averages.* On the technical side, our work introduces a *new methodological tool*: the average over block-structured random unitaries, representing a first step toward a systematic Weingarten calculus for unitaries with an energy structure. While we have shown how to treat block-diagonal Haar unitaries, the main open challenge is to extend this framework to genuinely smooth banded unitaries, such as the one shown in Fig. 2(c). Beyond the present setting, we expect this methodology to be of independent interest and to find applications in other communities where structured random unitaries naturally arise, for instance in the study of quantum dynamics with conserved quantities.

ACKNOWLEDGMENTS

We are grateful to R. Speicher and A. Wendel for their valuable and insightful feedback on free probability. We thank J. Kurchan for continuous collaborations on this topic. E.V. and S.P. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research

Foundation) under Germany’s Excellence Strategy—Cluster of Excellence Matter and Light for Quantum Computing (ML4Q) EXC 2004/1 -390534769, and DFG Collaborative Research Center (CRC) 183 Project No. 277101999—project B02. L.F. acknowledges support by the French government through the France 2030 program (PhOM—Graduate School of Physics), under reference ANR-11-IDEX-0003 (Project Mascotte, L. Foini).

DATA AVAILABILITY

The data that support the findings of this article are openly available [95].

APPENDIX A: MORE ON FREE PROBABILITY

In Sec. II A we introduced the basic notions of free probability theory, with particular emphasis on the combinatorial structure of partitions. In this section, we instead follow a more standard mathematical presentation, highlighting directly why free probability provides a natural framework for studying quantum operators in chaotic many-body quantum systems and their statistical properties. We then introduce operator-valued free probability, since some of the quantities introduced in this work can be naturally identified with objects already appearing in this framework. Finally, we discuss the isomorphism between partitions and permutations in more detail, as it will be essential for the derivations in the following appendices.

1. Standard introduction to free probability

a. *Non-commutative probability spaces.* The fundamental object in probability theory is a probability space $(\mathcal{X}, \varphi)$, consisting of a unital $*$ -algebra $\mathcal{X}$ and a linear functional $\varphi: \mathcal{X} \rightarrow \mathbb{C}$ satisfying

$$\varphi(1) = 1, \quad \varphi(x^\dagger x) \geq 0, \quad \varphi(x^\dagger) = \varphi^*(x).$$

When $\mathcal{X}$ is commutative, one recovers classical probability theory. A standard example is $\mathcal{X} = L^\infty(\Omega, P)$, the algebra of bounded random variables on a probability space (Ω, P) , with expectation functional $\varphi(x) = \int_\Omega x(\omega) dP(\omega)$.

Free probability extends this framework by allowing $\mathcal{X}$ to be non-commutative. Two important examples are: (i) the algebra $M_D(\mathbb{C})$ of $D \times D$ matrices equipped with the normalized trace $\varphi(x) = \text{Tr}(x)/D$; (ii) a von Neumann algebra of observables acting on a Hilbert space $\mathcal{H}$, together with a vector state $\varphi_\psi(x) = \langle \psi | x | \psi \rangle$. These two examples are closely connected. Indeed, free probability originated in the study of operator algebras, and, since large random matrices often provide finite-dimensional models for operator-algebraic structures, it was strongly influenced by random matrix theory; subsequently it supplied powerful tools for instance for studying the asymptotic eigenvalue distributions of large random matrices.

In the main text (and in the following sections), for a generic quantum observable A seen as a $D \times D$ matrix, we adopt the convention $\langle A \rangle \equiv \varphi(A) := \text{Tr}(A)/D$.

When $\mathcal{X}$ is non-commutative, many classical probabilistic notions require a nontrivial reformulation. The most important example is the concept of independence. For classically independent random variables x and y , mixed moments factorize. For instance,

$$\varphi(xyxy) = \varphi(x^2y^2) = \varphi(x^2)\varphi(y^2).$$

The question is how this property should be generalized in a non-commutative setting.

b. Freeness. The non-commutative analogue of independence is free independence, or *freeness*. Let X and Y be unital subalgebras of $\mathcal{X}$. They are said to be freely independent if

$$\varphi(c_1 c_2 \cdots c_n) = 0 \quad (\text{A1})$$

whenever:

- (i) each c_k belongs either to X or to Y ;
- (ii) consecutive elements come from different subalgebras;
- (iii) $\varphi(c_k) = 0$ for all k .

As an example, if $x \in X$ and $y \in Y$ are free, then

$$\varphi(xyxy) = \varphi(x^2)\varphi(y)^2 + \varphi(x)^2\varphi(y^2)\varphi(x)^2\varphi(y)^2.$$

Freeness and classical independence are distinct notions. In general, freely independent variables are not classically independent, and conversely.

A fundamental source of free random variables arises in random matrix theory: independent random matrices from ensembles such as the GUE become asymptotically free in the limit of large matrix dimension.

c. Further developments. Many central constructions of classical probability admit free analogs, including the free central limit theorem, free convolution, free cumulants, and their associated generating functions. A distinctive feature of free probability is its deep connection with combinatorial structures, most notably non-crossing partitions, as highlighted in Sec. II A.

2. Operator-valued free probability

Operator-valued free probability generalizes standard free probability by replacing the scalar-valued expectation $\varphi: \mathcal{X} \rightarrow \mathbb{C}$ with a *conditional* expectation $\varphi_{\mathcal{D}}: \mathcal{X} \rightarrow \mathcal{D}$ where $\mathcal{D}$ is a distinguished subalgebra of $\mathcal{X}$, satisfying the condition $\varphi_{\mathcal{D}}(\hat{d}_1 x \hat{d}_2) = \hat{d}_1 \varphi_{\mathcal{D}}(x) \hat{d}_2$, for $\hat{d}_{1,2} \in \mathcal{D}$. Moments retain information encoded in the chosen subalgebra, rather than collapsing all variables to scalar quantities: the choice of $\varphi_{\mathcal{D}}$ determines which degrees of freedom are averaged over and which are preserved. This additional structure makes operator-valued free probability

particularly well suited for systems with internal degrees of freedom.

In the context of random matrices, this framework is particularly useful for describing ensembles with non-trivial internal structure, so that the resulting moments preserve information about the underlying structure that would be lost under a global trace. We call *operator-valued moments* the matrices $\langle A \hat{d}_1 A \hat{d}_2 A \cdots \hat{d}_n A \rangle_{\mathcal{D}} \in \mathcal{D}$, with $\hat{d}_k \in \mathcal{D}$, $\forall k$. The corresponding *operator-valued free cumulants* are defined implicitly through the same moments-cumulants relation as in Eq. (17), that is

$$\langle A \hat{d}_1 A \cdots \hat{d}_n A \rangle_{\mathcal{D}} = \sum_{\mu \in NC(n+1)} \kappa_{\mu}^{\mathcal{D}}(A \hat{d}_1, \dots, A \hat{d}_n, A). \quad (\text{A2})$$

The difference with respect to scalar free probability is that the matrices $K_{\mu}^{\mathcal{D}}$ are not factorized in a product of block cumulants as in (17). Instead, the cumulants associated with the blocks of μ are recursively composed according to the nesting structure of the non-crossing partition, since they take values in $\mathcal{D}$.

For the purposes of this work, we restrict our attention to a specific example, that is the one of block matrices with the convenient choice of $\mathcal{D}$ given by a block-diagonal matrix with diagonal blocks. We focus on this particular setting because it admits a direct connection with the results presented in this work (see Appendix D).

Concretely, let us consider four blocks:

$$\begin{aligned} A &= \begin{pmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{pmatrix}, \quad \langle A \rangle_{\mathcal{D}} := \begin{pmatrix} \langle \mathcal{A}_{11} \rangle_{\mathbb{I}} & 0 \\ 0 & \langle \mathcal{A}_{22} \rangle_{\mathbb{I}} \end{pmatrix}, \\ \mathcal{D} &= \left\{ \begin{pmatrix} d_1 \mathbb{I} & 0 \\ 0 & d_2 \mathbb{I} \end{pmatrix} \right\}_{d_1, d_2 \in \mathbb{R}} \\ &= \text{span} \left\{ \hat{\Pi}_1 = \begin{pmatrix} \mathbb{I} & 0 \\ 0 & 0 \end{pmatrix}, \hat{\Pi}_2 = \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{I} \end{pmatrix} \right\} \end{aligned} \quad (\text{A3})$$

Using Eq. (A2), we have

$$\kappa_1^{\mathcal{D}}(A) = \text{diag}(\langle \mathcal{A}_{11} \rangle_{\mathbb{I}}, \langle \mathcal{A}_{22} \rangle_{\mathbb{I}}). \quad (\text{A4})$$

We can then calculate some examples of second-order operator-valued moments and free cumulants. We are interested in the ones weighted with the projectors on the diagonal blocks $\hat{\Pi}_{I=1,2}$:

$$\begin{aligned} \langle A \hat{\Pi}_1 A \rangle_{\mathcal{D}} &= \text{diag}(\langle \mathcal{A}_{11}^2 \rangle_{\mathbb{I}}, \langle \mathcal{A}_{21} \mathcal{A}_{12} \rangle_{\mathbb{I}}), \\ \langle A \hat{\Pi}_2 A \rangle_{\mathcal{D}} &= \text{diag}(\langle \mathcal{A}_{12} \mathcal{A}_{21} \rangle_{\mathbb{I}}, \langle \mathcal{A}_{22}^2 \rangle_{\mathbb{I}}). \end{aligned} \quad (\text{A5})$$

The moments-cumulants formula reads

$$\langle A \hat{\Pi}_I A \rangle_{\mathcal{D}} = \kappa_2^{\mathcal{D}}(A \hat{\Pi}_I, A) + \kappa_1(A \hat{\Pi}_I) \kappa_1(A), \quad (\text{A6})$$

and one can check that $\kappa_1(A \hat{\Pi}_1) = \kappa_1(A) \hat{\Pi}_1$. Therefore,

$$\begin{aligned}\kappa_2^{\mathcal{D}}(\hat{A}\hat{\Pi}_1, A) &= \text{diag}((\langle \mathcal{A}_{11}^2 \rangle - \langle \mathcal{A}_{11} \rangle^2) \mathbb{I}, \langle \mathcal{A}_{21} \mathcal{A}_{12} \rangle \mathbb{I}), \\ \kappa_2^{\mathcal{D}}(\hat{A}\hat{\Pi}_2, A) &= \text{diag}(\langle \mathcal{A}_{12} \mathcal{A}_{21} \rangle \mathbb{I}, (\langle \mathcal{A}_{22}^2 \rangle - \langle \mathcal{A}_{22} \rangle^2) \mathbb{I}).\end{aligned}\quad (\text{A7})$$

This can be generalized to higher order and for a generic number of blocks.

3. More on the isomorphism between partitions and permutations

In Sec. II A we briefly mentioned the isomorphism between partitions and permutations, without entering into its details. Here we clarify and formalize this correspondence, as it will be essential for the derivation of Eq. (19) in Appendix B.

a. Permutations. A permutation of n elements is a bijective rearrangement of the set $\{1, \dots, n\}$. Using the cyclic notation, we represent it as a sequence of disjoint cycles $\alpha = (C_1)(C_2) \cdots (C_{\#\alpha})$, where $\#\alpha$ is the number of cycles and each cycle (C_m) lists the elements that are mapped cyclically. Two special permutations are the identity e (n cycles of 1 element) and the maximal cycle $\gamma_n = (1\,2\,\dots\,n)$ (a single normal-ordered n -cycle).

b. Lattice. A useful notion is the one of rank or Cayley weight of a permutation $\alpha \in S_n$: $|\alpha| = n - \#\alpha$. The rank represents the minimal number of transpositions needed to generate α : it is minimized by e (rank 0) and maximized by γ_n (rank $n - 1$). The Cayley weight induces a natural metric on the symmetric group S_n ,

$$d(\alpha, \beta) = |\alpha^{-1}\beta| = n - \#(\alpha^{-1}\beta), \quad (\text{A8})$$

known as the Cayley distance. This distance measures the minimal number of transpositions required to transform α into β . With this metric, we consider the set of all permutations η that saturate the triangle inequality between α and β , that is, all permutations lying on some shortest path between α and β :

$$[\alpha, \beta] = \{\eta \in S_n : d(\alpha, \beta) = d(\alpha, \eta) + d(\eta, \beta)\}. \quad (\text{A9})$$

Between α and β there may exist multiple distinct shortest paths, corresponding to different sequences of transpositions, each defining a geodesic sequence

$$\alpha = \eta_0, \eta_1, \dots, \eta_k = \beta,$$

where each consecutive pair differs by a single transposition, $d(\eta_{i-1}, \eta_i) = 1$, and the total number of steps is minimal, $k = d(\alpha, \beta)$. This metric naturally defines the Cayley graph of S_n , whose vertices are all permutations and whose edges connect permutations differing by a single transposition. In this graph, the Cayley distance corresponds to the length of the shortest path between two vertices. Geodesics in the Cayley graph are therefore sequences of permutations connected by transpositions

that realize this minimal distance. Moreover, this notion of geodesic naturally induces a partial order on the set of permutations lying along geodesics between two fixed endpoints. Permutations on different single geodesics paths are, in general, not comparable under this order. However, restricted to a single geodesic, for instance one starting at the identity e , an order is defined as

$$\beta \leq \alpha \quad \text{if and only if } d(e, \beta) \leq d(e, \alpha), \quad (\text{A10})$$

that is, β lies closer to the identity than α , along the same geodesic. Under this order, permutations along geodesics between fixed endpoints form a lattice which is a subset of the Cayley graph [110]. A special role is played by the lattice of permutations in $[e, \gamma_n]$, as we will see below.

c. Dual. One can define a natural dual by

$$\alpha^* = \alpha^{-1}\gamma_n. \quad (\text{A11})$$

While the dual can be defined for any permutation in S_n , it is an order-reversing bijection only when considering permutations lying along geodesics between two fixed endpoints. Therefore, if $\beta \leq \alpha$ along the geodesic, then $\beta^* \geq \alpha^*$. When considering permutations in $[e, \gamma_n]$, intuitively, the dual exchanges positions relative to the two endpoints, mapping permutations close to e to permutations close to γ_n and vice versa. If $\alpha \in [e, \gamma_n]$, then also $\alpha^* \in [e, \gamma_n]$, and $|\alpha| + |\alpha^*| = n - 1$.

d. Embedding partitions-permutations. As said in the main text, every permutation $\alpha \in S_n$ induces a partition μ_α by grouping together the elements belonging to the same cycle: $\mu_\alpha = \{\text{cycles of } \alpha\}$. The converse is not true: a partition does not specify a unique permutation, because one must choose an order of the elements within each block to define a cycle. However, when restricting the symmetric group to a geodesic interval from the identity to a maximal cycle, the correspondence becomes bijective. We focus in particular on $[e, \gamma_n]$; the cycles of a permutation can be represented as non-crossing blocks when the elements are arranged on a circle in the natural order, that is, when they lie on the interval $[e, \gamma_n]$. Conversely, any non-crossing partition can be uniquely realized by a permutation lying on $[e, \gamma_n]$, since the circular order of elements is fixed. Not only does this establish an isomorphism between the geodesic interval in the symmetric group and the lattice of non-crossing partitions, but it also preserves the partial order [111]: if one permutation is smaller than another along the geodesic in $[e, \gamma_n]$, the corresponding non-crossing partitions are ordered in the same way according to the refinement order. In this sense, the isomorphism respects the lattice structure. Moreover, it turns out that the dual permutation, when viewed as a partition, is the Kreweras complement. Therefore, we have the well known embedding of $NC(n)$ into S_n :

$$([e, \gamma_n], \leq_{\text{geod}}) \cong (NC(n), \leq_{\text{ref}}). \quad (\text{A12})$$

APPENDIX B: GLOBAL ROTATIONAL INVARIANCE DERIVATIONS

In this Appendix we derive Eq. (19) given in the main. The first step consists in inserting Eq. (12) into Eq. (11), meaning inserting the averages over the Haar ensemble in the expression for the matrix elements' moments. As a first step, we obtain:

$$\begin{aligned} \overline{A_{i_1 j_1}^U \cdots A_{i_n j_n}^U} &= \sum_{\alpha \in S_n} \sum_{\beta \in S_n} W_{\mathbf{g}_{\alpha, \beta}}(D) \sum_{\tilde{i}_1 \cdots \tilde{i}_n} A_{i_1 \tilde{i}_{\beta(1)}} \cdots A_{i_n \tilde{i}_{\beta(n)}} \delta_{i_{\alpha(1)} j_1} \cdots \delta_{i_{\alpha(n)} j_n} \\ &\equiv \sum_{\alpha \in S_n} \frac{\mathcal{K}_\alpha}{D^{n-\#\alpha}} \delta_{i_{\alpha(1)} j_1} \cdots \delta_{i_{\alpha(n)} j_n}, \end{aligned} \quad (\text{B1})$$

where we defined, with prefactor $D^{n-\#\alpha}$ for convenience:

$$\mathcal{K}_\alpha = D^{n-\#\alpha} \sum_{\beta \in S_n} W_{\mathbf{g}_{\alpha, \beta}}(D) \sum_{\tilde{i}_1 \cdots \tilde{i}_n} A_{i_1 \tilde{i}_{\beta(1)}} \cdots A_{i_n \tilde{i}_{\beta(n)}} = \sum_{\beta \in S_n} \frac{D^{n-\#\alpha}}{D^{-\#\beta}} W_{\mathbf{g}_{\alpha, \beta}}(D) \prod_{m=1}^{\#\beta} \langle A^{l_m} \rangle. \quad (\text{B2})$$

The same expression can also be understood as the action of the corresponding Haar quantum channel on $A^{\otimes n}$ in the energy basis, following the formulation of Ref. [112].

We now explain terms appearing in this expression:

(i) $W_{\mathbf{g}_{\alpha, \beta}}(D)$ is the Weingarten function, a rational function of the Hilbert space dimension D . Viewed as a matrix, it is defined as the pseudo-inverse of the Gram matrix with entries $G_{\alpha, \beta} = D^{(\alpha^{-1}\beta)}$. The Weingarten function depends only on the cycle structure of $\alpha^{-1}\beta$. A well known result is the first-order contribution of the Weingarten function when expanded for large D :

$$W_{\mathbf{g}_{\alpha, \beta}}(D) = \frac{1}{D^{2n-\#(\alpha^{-1}\beta)}} \left[\phi(\alpha, \beta) + \mathcal{O}\left(\frac{1}{D^2}\right) \right]. \quad (\text{B3})$$

$\phi(\alpha, \beta)$ is a known multiplicative function, it factorizes according to the cycles of $\alpha^{-1}\beta$ as $\prod_{m=1}^{(\alpha^{-1}\beta)} (-1)^{l_m-1} C_{l_m-1}$, with C_n Catalan number;

(ii) in the last equality we defined $\langle \bullet \rangle = \text{Tr}(\bullet)/D$ as the normalized trace, and we used the equality

$$\sum_{\tilde{i}_1 \cdots \tilde{i}_n} A_{i_1 \tilde{i}_{\beta(1)}} \cdots A_{i_n \tilde{i}_{\beta(n)}} = \prod_{m=1}^{\#\beta} \text{Tr}(A^{l_m}). \quad (\text{B4})$$

Similarly to the calculation in Refs. [43,113], done at the level of the Haar quantum channel, we perform a systematic expansion in powers of $1/D$ of Eq. (B2), for a fixed α : inserting the expansion of the Weingarten function [Eq. (B3)] we get

$$\mathcal{K}_\alpha = \sum_{\beta \in S_n} \frac{D^{n-\#\alpha}}{D^{\Delta(\alpha, \beta)}} \left[\phi(\alpha, \beta) \prod_{m=1}^{\#\beta} \langle A^{l_m} \rangle + \mathcal{O}\left(\frac{1}{D^2}\right) \right], \quad (\text{B5})$$

where we substituted $\Delta(\alpha, \beta) = 2n - (\alpha^{-1}\beta) - \#\beta = d(e, \beta) + d(\beta, \alpha)$ with $d(\cdot, \cdot)$ Cayley distance, defined in Eq. (A8).

The leading order is given only by the β s that minimize $\Delta(\alpha, \beta)$. This happens when $\beta \in [e, \alpha]$, so that $\Delta(\alpha, \beta) = n - \#\alpha$, its minimum value. Moreover, the condition $\beta \in [e, \alpha]$ is equivalent to having $\beta \leq \alpha$.

Permutations that do not belong to the geodesic interval give subleading contributions of order $\mathcal{O}(1/D^{n-\#\alpha+2})$ and so must be taken into account along the Weingarten approximation. Indeed, one can show [114] that $\nexists \alpha, \beta$ so that $\Delta(\alpha, \beta) = n - \#\alpha + 1$.

Therefore, the leading order is given by

$$\mathcal{K}_\alpha = \sum_{\beta \leq \alpha} \phi(\alpha, \beta) \prod_{m=1}^{\#\beta} \langle A^{l_m} \rangle + \mathcal{O}\left(\frac{1}{D^2}\right). \quad (\text{B6})$$

We can interpret the result obtained in terms of partitions instead of permutations. Specifically, the leading order of $\mathcal{K}_\alpha$ is the free cumulant κ_α associated with the non-crossing partition defined by α . In order to see the equivalence, we write explicitly the inverse relation of Eq. (17) which expresses a generic cumulant κ_μ , for $\mu \in NC(n)$ in terms of the moments, via the Möbius function ψ :

$$\kappa_\mu = \sum_{\substack{\nu \leq \mu \\ \nu \in NC(n)}} \psi(\nu, \mu) \prod_{m=1}^{\#\nu} \langle A^{l_m} \rangle. \quad (\text{B7})$$

Considering α and β as non-crossing partitions [115], the sum in Eq. (B6) runs over $\beta \leq \alpha, \beta \in NC(n)$, as in Eq. (B7). Moreover, one can identify ϕ with the Möbius function. It is now straightforward to see that $\kappa_\alpha = \kappa_\alpha + \mathcal{O}(1/D^2)$.

For the matrix elements' moments (B1) it follows

$$\overline{A_{i_1 j_1}^U \cdots A_{i_n j_n}^U} = \sum_{\alpha \in S_n} \frac{1}{D^{n-\#\alpha}} \cdot \left[\kappa_\alpha \delta_{i_{\alpha(1)} \cdot j_1} \cdots \delta_{i_{\alpha(n)} \cdot j_n} + \mathcal{O}\left(\frac{1}{D^2}\right) \right]. \quad (\text{B8})$$

This expression tells us that the matrix elements' moments are given as a sum over permutations, each coming with multiple δ s. In other words, a product of matrix elements, averaged on the ensemble, is non-zero only when the second set of indices $\{j_1 \cdots j_n\}$ is a permutation of the firsts $\{i_1 \cdots i_n\}$, or also, each index is repeated at least once in the product. This reflects the fact that the products in (12) are non-zero only when a U couples with a U^* . This is typical of the unitary group. As a consequence, the non-zero products of matrix elements can be written as

$$\overline{A_{i_1 i_{\sigma(1)}}^U \cdots A_{i_n i_{\sigma(n)}}^U} = \sum_{\alpha \in S_n} \frac{1}{D^{n-\#\alpha}} \left[\kappa_\alpha \delta_{\alpha\sigma}(\mathbf{i}) + \mathcal{O}\left(\frac{1}{D^2}\right) \right], \quad (\text{B9})$$

where we defined $\delta_{\alpha\sigma}(\mathbf{i}) = \delta_{i_{\alpha(1)}, i_{\sigma(1)}} \cdots \delta_{i_{\alpha(n)}, i_{\sigma(n)}}$ to keep the notation shorter. We will also use the notation $\mathbf{i} := \{i_1, \dots, i_n\}$ and $\sigma(\mathbf{i}) := \{i_{\sigma(1)}, \dots, i_{\sigma(n)}\}$. Differently from Eq. (19) and the surrounding discussion, here σ is a generic permutation, since we are still considering all the possible products of matrix elements; however, it is nontrivial to define the permutation σ , for a given product of matrix elements, since in $\mathbf{i}$ there could be repeated indices.

(i) When there are no repetitions, then σ is uniquely defined: there is a bijective correspondence between a product of matrix elements with different indices and a permutation. For instance, $\sigma = (12)(34)$ defines uniquely $A_{ij}A_{ji}A_{kl}A_{lk}$, for $i \neq j \neq k \neq l$.

(ii) When there are repetitions, instead, there are multiple permutations giving the same product of matrix elements; we define σ as the finest, that is, the one with the biggest number of cycles. For instance, one can use both (12)(34) or (1234) for defining $A_{ij}A_{ji}A_{il}A_{li}$, but we chose σ as the first. We notice that, in this way, each cycle of σ

involves only different indices in $\mathbf{i}$, meaning that $i_{C_m(1)} \neq \cdots \neq i_{C_m(l_m)}$, for all the cycles C_m of σ of length l_m . However, the finest permutation might not be unique: there could be different σ s with the same number of cycles. This is the case for $A_{ij}A_{ji}A_{ij}A_{ji}$ which can be defined by two different finest permutations (12)(34) and (14)(23), both with two cycles. This happens only for the product of matrix elements that were called non-cacti.

It is clear that a permutation σ will not uniquely define a product of matrix elements, since the same permutation could be describing a product with all different indices or with repeated indices, being the finest. This is the case, for instance, for $A_{ij}A_{ji}A_{ii}A_{ii}$ and $A_{ij}A_{ji}A_{ii}A_{jj}$, where $\sigma = (12)(3)(4)$. Therefore, together with σ , one needs to give the additional ordered set of indices $\mathbf{i}$ on which the permutation acts, with the constraint that σ is the finest, if there are repetitions. In the example, $\mathbf{i} = \{ijii\}$ for the first, and $\mathbf{i} = \{ijij\}$ for the second.

We highlight a special case that happens when $i_k = i_{\sigma(k-1)} \bmod n$. The condition depends both on σ and $\mathbf{i}$; of course is satisfied when $\sigma = \gamma_n$ and there are no repeated indices, but also, for instance, for $\sigma = (12)(3)(4)$ on $\mathbf{i} = \{ijii\}$. More specifically, this condition defines what we called ETH-cycles, for which the products of matrix elements are of the form $\overline{A_{i_1 i_2}^U A_{i_2 i_3}^U \cdots A_{i_n i_1}^U}$, where indices can be different or repeated. Excluding the non-cacti, as explained in the main, these products of matrix elements can be described uniquely by a non-crossing partition [116], whose Kreweras complement, seen as a permutation, is exactly σ ; moreover, $\sigma \in [e, \gamma_n]$ in those cases.

We will now focus the discussion on ETH-cycles, excluding the non-cacti, that is simple loops and cacti. For those, we can derive the closed formula Eq. (19), while the destiny of the others will be discussed separately afterwards.

1. Closed formula for simple loops and cacti

From expression (B9), we can deduce both the leading and the first subleading contributions.

The $\delta_{\alpha\sigma}(\mathbf{i})$ appearing in the equation is selecting only some α s in the sum over the symmetric group, which are the ones that act as σ on $\mathbf{i}$. One could initially assume that the condition is satisfied only when $\alpha = \sigma$; this is indeed correct if σ acts on a set of distinct indices, corresponding to the case of simple loops. For cacti, instead, α s are all the permutations giving the same result as σ , when applied to the set with repeated indices. As said above, with repeated indices σ is not well defined, but we chose it to be the finest. Permutations α s are all the others which could also define the same product of matrix elements, but have more cycles.

a. Leading order

Between all the α s satisfying the $\delta_{\alpha\sigma}(\mathbf{i})$, the leading order is given by the one minimizing $n - \#\alpha$, the exponent of $1/D$. This happens for the permutation with the biggest number of cycles i.e. the finest which is $\alpha = \sigma$, for how we defined it. Therefore,

$$\overline{A_{i_1 i_{\sigma(1)}}^U \cdots A_{i_n i_{\sigma(n)}}^U} = \frac{\kappa_\sigma}{D^{n-\#\sigma}} + \mathcal{O}\left(\frac{1}{D^{n-\#\sigma+1}}\right), \quad (\text{B10})$$

with

$$\mathcal{O}\left(\frac{1}{D^{n-\#\sigma+1}}\right) = \sum_{\alpha \neq \sigma \in S_n} \frac{1}{D^{n-\#\alpha}} \left[\kappa_\alpha \delta_{\alpha\sigma}(\mathbf{i}) + \mathcal{O}\left(\frac{1}{D^2}\right) \right] + \mathcal{O}\left(\frac{1}{D^{n-\#\sigma+2}}\right). \quad (\text{B11})$$

For simple loops and cacti we identify σ with π , and we obtain the expression in the main Eq. (19). In particular, for simple loops, $\pi = \gamma_n$; as a consequence, they are the smallest product of matrix elements, having the scaling $1/D^{n-1}$.

b. First subleading correction

We want to read the first subleading term from (B11). We immediately see that this is given by all the α s satisfying the $\delta_{\alpha\sigma}(\mathbf{i})$ for which $\#\alpha = \#\sigma - 1$, meaning the ones with one cycle less than σ . For convenience, we call them ρ s. Therefore, from (B11):

$$\mathcal{O}\left(\frac{1}{D^{n-\#\sigma+1}}\right) = \sum_{\substack{\rho \in S_n \text{ with} \\ \#\rho = \#\sigma - 1}} \frac{\kappa_\rho}{D^{n-\#\rho}} \delta_{\rho\sigma}(\mathbf{i}) + \mathcal{O}\left(\frac{1}{D^{n-\#\sigma+2}}\right). \quad (\text{B12})$$

We emphasize that ρ s are not forced to be in the geodesic interval $[e, \gamma_n]$, even if σ is. Before providing a geometric interpretation of the ρ s that appear in the expansion, we remark that only the first subleading correction admits a compact and explicit expression. Higher-order contributions require taking into account not only permutations α with progressively fewer cycles, but also subleading terms in $\mathcal{K}_\alpha$ arising from the higher-order structure of the Weingarten function and from permutations that do not belong to the geodesic, as indicated in Eq. (B5). Accordingly, contributions of order $\mathcal{O}(1/D^{n-\#\sigma+2})$ lie beyond the analysis of this paper [117].

Let's then focus on the ρ s. Firstly, we already stressed that for simple loops there are no other permutations acting as σ , apart from σ itself, since in $\mathbf{i}$ there are no repetitions. In this case, the first-order correction is zero. Since the leading order has scaling $1/D^{n-1}$, the error is of order $\mathcal{O}(1/D^{n+1})$.

When indices are repeated, we want to show that the condition given by $\delta_{\rho\sigma}(\mathbf{i})$ restricts the sum for $\rho \in S_n$, with the additional condition $\#\rho = \#\sigma - 1$, only over $\rho \in [e, \gamma_n]$.

We introduce the following two notations:

- (i) $\lambda := \sigma\rho^{-1}$. The permutation λ permits to obtain σ when composed with ρ . The $\delta_{\rho\sigma}(\mathbf{i})$ implies $i_k = i_{\lambda(k)} \forall k$, that is λ acting as the identity on $\mathbf{i}$;
- (ii) $\hat{\alpha} := \gamma_n \alpha^{-1}$ for a generic permutation $\alpha \in S_n$, which is a sort of anti-dual, with respect to Eq. (A11). Indeed, it is still an order-reversing bijection; however it inverts the usual direction, bringing a dual permutation to the original one. For instance for permutations on the geodesic interval $[e, \gamma_n]$, one can see α as the Kreweras complement of $\hat{\alpha}$. As a consequence, $\hat{\sigma}$ is basically the non-crossing partition characterizing the repeated indices in $\mathbf{i}$. Therefore, we will say that elements in the same cycle of $\hat{\sigma}$ are associated with equal indices in $\mathbf{i}$, meaning that $i_{\hat{\sigma}_m(1)} = \cdots = i_{\hat{\sigma}_m(l_m)}$, for all the cycles $\hat{\sigma}_m$ of $\hat{\sigma}$ of length l_m .

The statement we want to prove is the following: under the assumptions (1) σ satisfies the condition of being cyclic on the set $\mathbf{i}$, $i_k = i_{\sigma(k-1)} \bmod n$, and (2) $\#\rho = \#\sigma - 1$, then

$$\{\rho \in S_n / i_{\sigma(k)} = i_{\rho(k)} \forall k\} \iff \rho \in [e, \gamma_n], \quad (\text{B13})$$

that is, all and only the ρ s satisfying $\delta_{\rho\sigma}(\mathbf{i})$ are the ones on the geodesic interval $[e, \gamma_n]$.

We start by showing the left to right implication; as a first step, we show that if ρ satisfies $\delta_{\rho\sigma}(\mathbf{i})$ and $\#\rho = \#\sigma - 1$, then $d(\rho, \sigma) = 1$, that is

$$\#(\rho^{-1}\sigma) \equiv \#(\sigma\rho^{-1}) \equiv \#\lambda = n - 1.$$

The reader should notice that the condition (2) is equivalent to $d(e, \rho) = d(e, \sigma) + 1$, but does not imply directly that $\rho > \sigma$ which means $\sigma \in [e, \rho]$. This happens when the additional condition $d(\rho, \sigma) = 1$ is satisfied. We want to show that λ is a transposition; the permutation λ needs to satisfy two conditions: it must unify two cycles of σ which are distinct, so that ρ has one cycle less than σ , and it must act as the identity on the set of indices $\mathbf{i}$. This means that there should be a cycle of λ , let's say C , with elements belonging to two different cycles of σ (not more than two), and so that $i_{C(1)} = \cdots = i_{C(l)}$. Since in a single cycle of σ there cannot be elements associated with the same index, by construction of σ , the cycle C of λ cannot contain more than two elements, that is $l = 2$. We will label these elements as x and y . Moreover, there can be only one cycle containing more than one element, all the other cycles of λ need to have unit length. Let us admit this is not the case: let us imagine we have more than one cycle having more than one element. This could happen when more than two cycles of σ are involved and so unified, which goes against $\#\rho = \#\sigma - 1$, or when σ is one of the permutations defining

a non-cacti. Indeed, these are the only cases in which different cycles could be associated with more than one equal index, meaning that in two different cycles of σ , saying C_1 and C_2 , more than one couple (m, n) exists, so that $i_{C_1(m)} = i_{C_2(n)}$. These cases are excluded from this derivation, since we consider only simple loops and cacti. Therefore, λ is a transposition, and $\rho > \sigma$ with one cycle less.

As a successive step, one can directly show that $\sigma\rho^{-1} = \hat{\sigma}^{-1}\hat{\rho}$, which implies $d(\hat{\rho}, \hat{\sigma}) = 1$, for what we just showed. If two permutations are at distance one, they must have one cycle of difference, that is

$$d(\hat{\rho}, \hat{\sigma}) = 1 \Rightarrow \#\hat{\rho} - \#\hat{\sigma} = \pm 1 : \quad (\text{B14})$$

$\hat{\rho}$ has one more cycle than $\hat{\sigma}$ or one less.

We also notice that, as a consequence of their definitions, $\hat{\sigma}^{-1}\hat{\rho} = \lambda$, which means that the same transposition λ of above, composed with $\hat{\sigma}$ gives $\hat{\rho}$. Therefore, λ is or dividing a cycle of $\hat{\sigma}$ in two cycles of $\hat{\rho}$ [+ in Eq. (B14)], or vice versa unifying two cycles of $\hat{\sigma}$ in one of $\hat{\rho}$ [- in Eq. (B14)]. To decide the right direction, we have to look at the elements $x, y \in C$ and understand if they are in a same cycle of $\hat{\sigma}$ and different of $\hat{\rho}$, or vice versa; for construction, $i_x = i_y$, that is, x and y are constrained to be in the same cycle of $\hat{\sigma}$. Therefore, λ is dividing the cycle of $\hat{\sigma}$ containing both x and y in two different cycles of $\hat{\rho}$, one with x and another with y . As a consequence, $\#\hat{\rho} = \#\hat{\sigma} + 1$ and $\hat{\rho} < \hat{\sigma}$.

Since $\sigma \in [e, \gamma_n]$, then also $\hat{\sigma} \in [e, \gamma_n]$, and since $\hat{\rho} < \hat{\sigma}$, the same is true for ρ . We notice that $\hat{\rho} < \hat{\sigma}$ is a stronger condition than $\rho > \sigma$, because it implies that $\hat{\rho} \in [e, \gamma_n] \Rightarrow \rho \in [e, \gamma_n]$. Indeed, there are ρ 's bigger than σ of one cycle, which do not satisfy the delta function, which are indeed all the ones not belonging to $[e, \gamma_n]$.

Let us now show the right to left implication in Eq. (B13); when $\rho > \sigma$, $\in [e, \gamma_n]$ with one cycle less, then $\hat{\rho} < \hat{\sigma}$, $\in [e, \gamma_n]$ with one cycle more. This means that each cycle of $\hat{\rho}$ is included in a cycle of $\hat{\sigma}$. Specifically, considering a generic cycle $\hat{C}$ of $\hat{\sigma}$ of length l , it can be written as the disjoint union of all the cycles of $\hat{\rho}$ contained in $\hat{C}$, that we indicate as $\hat{\mathcal{R}}_g$, for $g = 1 \cdots r$, admitting they are r . Since it is true that $i_{\hat{C}(1)} = \cdots = i_{\hat{C}(l)}$, it is also true that $i_{\hat{\mathcal{R}}_1(1)} = \cdots = i_{\hat{\mathcal{R}}_1(l_1)} = i_{\hat{\mathcal{R}}_2(1)} \cdots = i_{\hat{\mathcal{R}}_2(l_2)} = \cdots = i_{\hat{\mathcal{R}}_g(1)} = \cdots = i_{\hat{\mathcal{R}}_g(l_g)}$. Not only, this would be true even if instead of $\hat{\sigma}$ and $\hat{\rho}$ we consider their inverses, since the inverse operation is not mixing cycles, but just changing the order of the elements in the cycles. This is valid for each cycle of $\hat{\sigma}$, therefore it is clear that $\forall k$:

$$\begin{aligned} i_{\hat{\rho}(k)} = i_{\hat{\sigma}(k)} &\Leftrightarrow i_{\hat{\rho}^{-1}(k)} = i_{\hat{\sigma}^{-1}(k)} \Leftrightarrow i_{\hat{\sigma}^{-1}\hat{\rho}(k)} = i_k \\ &\Leftrightarrow i_{\sigma\rho^{-1}(k)} = i_k \Leftrightarrow i_{\sigma(k)} = i_{\rho(k)}, \end{aligned} \quad (\text{B15})$$

same condition imposed from the delta function $\delta_{\rho\sigma}(\mathbf{i})$, as we wanted to prove.

This completes the proof.

An important remark is that the condition of $\alpha \in [e, \gamma_n]$ would not be necessary and sufficient for satisfying $\delta_{\alpha\sigma}(\mathbf{i})$ when the assumption (2) is made for higher order corrections, that is, looking for a generic order M for which $\#\alpha = \#\sigma - M$. This would still imply $d(\alpha, \sigma) = d(\hat{\alpha}, \hat{\sigma}) = M$, but it would not be sufficient for saying that $\#\hat{\rho} - \#\hat{\sigma}$ is equal to $\pm M$; the implication (B14) is true only for $M = 1$. Therefore, in all the other cases, taking only $\alpha \in [e, \gamma_n]$ would be too restrictive: there would be permutations α satisfying $\delta_{\alpha\sigma}(\mathbf{i})$, but not on the geodesic interval. Since we already noticed that the higher corrections would be much more difficult to interpret, we do not go further in this direction.

Example. We consider useful to follow the derivation with a specific example. We choose $\mathbf{i} = \{ijji\}$ and $\sigma = (12)(3)(4)$, that is the ETH-cycle $\overline{A_{ij}^U A_{ji}^U A_{ii}^U A_{ii}^U}$. We construct the permutations λ 's so that they unify one single cycle of σ , with the constraint that elements in the same cycle are associated with equal indices. There are only three choices, and they are all constrained to be transpositions; with the respective $\rho = \lambda\sigma$, they are $\lambda_1 = (13)(2)(4) \rightarrow \rho_1 = (123)(4)$, $\lambda_2 = (14)(2)(3) \rightarrow \rho_2 = (124)(3)$, and $\lambda_3 = (34)(1)(2) \rightarrow \rho_3 = (12)(34)$. Clearly, for all of them $d(\rho, \sigma) = 1$. However, we notice that there were other permutations with one fewer cycle, but not at distance one, for instance $(142)(3)$.

Calculating the anti-dual permutation, $\hat{\sigma} = (134)(2)$ and $\hat{\rho}_1 = (14)(2)(3)$, $\hat{\rho}_2 = (34)(1)(2)$, $\hat{\rho}_3 = (13)(2)(4)$. We notice that $d(\hat{\rho}, \hat{\sigma}) = 1$ in all the cases, and moreover $\#\hat{\rho} = \#\hat{\sigma} + 1$ which is implied by the fact that the elements exchanged by the transpositions λ are always belonging to the same cycle of $\hat{\sigma}$. Moreover, the relation $\hat{\rho} = \hat{\sigma}\lambda$ is valid. The ρ 's satisfying the delta function, as reported in the main in Eq. (20c), are all and only the ones in the geodesic interval $[e, \gamma_4]$.

Let us discuss quickly which permutations would be involved for $M = 2$; in this case the permutations satisfying the delta function are $\alpha_1 = (1234)$ and $\alpha_2 = (2134) \notin [e, \gamma_4]$. It is still true that $d(\alpha, \sigma) = 2$ (even if it is evident that not all the permutations at distance 2 satisfy the delta function), however, in the dual the permutations are not forced to have two cycles more than $\hat{\sigma}$, for instance $\hat{\alpha}_2 = (132)(4)$ has the same number of cycles of $\hat{\sigma}$.

Therefore we proved that Eq. (B12) is equivalent to:

$$\mathcal{O}\left(\frac{1}{D^{n-\#\sigma+1}}\right) = \sum_{\substack{\rho \in [e, \gamma_n] \text{ with} \\ \rho > \sigma, \#\rho = \#\sigma - 1}} + \mathcal{O}\left(\frac{1}{D^{n-\#\sigma+2}}\right). \quad (\text{B16})$$

We remind the reader of the assumption that σ is defining only ETH-cycles, excluding the non-cacti. In

this case we can finally translate the formula to the language of partitions. Summing over $\rho \in [e, \gamma_n] > \sigma$ at distance 1 is equivalent to summing over the non-crossing partitions associated, at distance 1; in other words, over $\tau \in NC(n) > \pi$ at distance 1. Therefore, Eqs. (B10), (B16) give Eq. (19) in the main. Some examples were presented in Eqs. (20a)–(20d).

4. Destiny of non-ETH-cycles and non-cacti

Equation (B9) is valid for all the products of matrix elements, and $\forall \sigma \in S_n$, however, Eqs. (B10) and (B16) were derived only for simple loops and cacti, that is, when $\sigma \in [e, \gamma_n]$, and it is cyclic on the set $\mathbf{i}$. Let us examine where the derivation fails for the other cases of non-ETH-cycles and non-cacti.

a. Non-cacti

We said above that non-cacti cannot be described by a unique permutation σ . Indeed, the finest permutation is not unique: different σ 's with the same number of cycles give the matrix product of non-cacti. As a consequence, the leading order will not be given by a single permutation, but by all the finest σ 's, summed together. This reflects the non-existence of the Kreweras complement for the crossing partition and the fact that a crossing diagram cannot be factorized uniquely, as instead happens for all the other ETH-cycles. For instance, the unique non-cacti of $n = 4$ is given by

$$\overline{A_{ij}^U A_{ji}^U A_{ij}^U A_{ji}^U} = \frac{2\kappa_2^2}{D^2} + \mathcal{O}\left(\frac{1}{D^3}\right). \quad (\text{B17})$$

Since our interest regards simple loops and cacti, we don't discuss further the subleading of the non-cacti.

b. Non-ETH-cycles

Now we discuss the case of products of matrix elements which are not of the form $A_{i_1 i_2}^U A_{i_2 i_3}^U \cdots A_{i_n i_1}^U$, that is, they do not satisfy $i_k = i_{\sigma(k-1)} \bmod n$, and for which σ is uniquely defined (we exclude all the cases similar to the cacti). The condition depends both on σ and $\mathbf{i}$.

When there are no repeated indices in $\mathbf{i}$, whether σ is cyclic or not, the leading order is given by Eq. (B10), and the subleading is zero. This was the case for simple loops and holds in general for all the permutations defined on $\mathbf{i}$ without repetitions, since no other permutations are satisfying $\delta_{\sigma\sigma}(\mathbf{i})$.

When there are repeated indices in $\mathbf{i}$, the leading order is still given by Eq. (B10), because one can still define a unique finest permutation; this contribution will be the same as the one of the ETH-cycles with the same σ , but different $\mathbf{i}$. The subleading, instead, is still given by Eq. (B12), but not by Eq. (B16). Indeed, its derivation relies on the assumption of σ being cyclic on $\mathbf{i}$ and breaks

down when this is not true. In particular what is important in the proof is that, when σ has this property, $\hat{\sigma}$ is basically the non-crossing partition characterizing the repeated indices in $\mathbf{i}$, so that, $i_{\hat{\sigma}_m(1)} = \cdots = i_{\hat{\sigma}_m(l_m)}$, for all the cycles $\hat{\sigma}_m$ of $\hat{\sigma}$ of length l_m . This was essential for proving that Eq. (B14) is satisfied with the sign $+$; for non-ETH-cycles, x, y , elements of the transposition λ , are not forced to be in the same cycle of $\hat{\sigma}$, therefore the implication can be valid both for $\pm$. As a consequence, the condition $\hat{\rho} < \hat{\sigma}$ is not implied and can happen that $\rho \notin [e, \gamma_n]$ still satisfies the $\delta_{\rho\sigma}(\mathbf{i})$.

As an example, let us consider $\overline{A_{ij}^U A_{ji}^U A_{ii}^U A_{jj}^U}$, in order to compare it with $\overline{A_{ij}^U A_{ji}^U A_{ii}^U A_{ii}^U}$ that we discussed above. First of all, $\sigma = (12)(3)(4)$ in both the cases and the leading order is exactly the same for both averages. Focusing on the non-ETH cycle, we construct the permutations λ s so that they unify one single cycle of σ , with the constraint that elements in the same cycle are associated with equal indices. There are only two choices, and they are all constrained to be transpositions; with the respective $\rho = \lambda\sigma$, they are $\lambda_1 = (13)(2)(4) \rightarrow \rho_1 = (123)(4)$, and $\lambda_2 = (24)(1)(3) \rightarrow \rho_2 = (142)(3)$. Clearly, for all of them $d(\rho, \sigma) = 1$. Indeed, this part of the proof relied only on the definition of σ being the finest. Problems come when calculating the anti-dual permutations, $\hat{\rho}_1 = (14)(2)(3), \hat{\rho}_2 = (1342)$. We still have $d(\hat{\rho}, \hat{\sigma}) = 1$ in both cases, but $\#\hat{\rho}_2 = \#\hat{\sigma} - 1$. Indeed the elements exchanged by the transposition λ_2 do not belong to the same cycle of $\hat{\sigma}$, since $\hat{\sigma} = (123)(4)$ is mixing the positions of equal indices.

Again, since our interest regards the simple loops and cacti, we don't discuss further the subleading corrections of non-ETH-cycles.

To summarize, the leading order given as Eq. (B10) is valid for all the matrix elements' moments apart from when σ cannot be uniquely defined. Equation (B16), instead, is not valid for non-ETH-cycles with repeated indices, and, again, for the non-cacti.

APPENDIX C: LOCAL ROTATIONAL INVARIANCE DERIVATIONS

In this Appendix we evaluate Eq. (16). In contrast to the globally invariant case, the matrices here behave as random unitaries only within their respective index subsets. Consequently, Eq. (12) and Weingarten calculus cannot be applied straightforwardly. We explicitly evaluate the first few n orders, while showing that the Weingarten formalism can still be employed by properly coupling matrices acting on different intervals. Using the same approximations as in the global case will lead to similar results, now encoding the energy dependencies. For the computation, we employ the minimal block model in Fig. 2(b), and we show afterward how to translate the expressions to the more refined model in Fig. 2(c). We will

consider, since the beginning, ETH-cycles only, starting from simple loops in order to derive Eqs. (21), (23) given in the main, and then evaluating explicitly some examples of cacti in order to show how the factorization and the first subleading correction appear in the case of rotational invariance.

1. $n = 1$

Since $U^{(I)}$ and $U^{(J)}$ have two independent unitaries, respectively only to the set of indices I and J , the average in Eq. (C1) is non-zero only when $I = J$:

$$\overline{A_{ij}^U} = \sum_{\substack{\bar{i} \in I \\ \bar{j} \in J}} \overline{U_{ii}^{(I)*} U_{jj}^{(J)}} A_{i\bar{j}} = \delta_{I,J} \sum_{\bar{i}, \bar{j} \in I} \overline{U_{ii}^{(I)*} U_{jj}^{(I)}} A_{i\bar{j}}. \quad (\text{C1})$$

After fixing this condition, the product is a rotation that involves only the submatrix $A_{\bar{i}\bar{j}}$ for $\bar{i}, \bar{j} \in I$; then, the average can be calculated through the Weingarten calculus for the dimension of the submatrix, that is D_i . From Eq. (12):

$$\overline{U_{ii}^{(I)*} U_{jj}^{(I)}} = \text{Wg}_{e,e}(D_i) \delta_{ij} \delta_{\bar{i}\bar{j}}.$$

Therefore, since δ_{ij} directly implies $\delta_{I,J}$,

$$\overline{A_{ij}^U} = \delta_{ij} \frac{1}{D_i} \sum_{\bar{i} \in I} A_{i\bar{i}} \equiv \delta_{ij} K_1(E_i),$$

where $K_1(E_i)$ depends on the mesoscopic energy E_i of the interval I .

It is straightforward to translate this result to the smoothed version, just explicitly writing

$$K_1(E_i) = \frac{1}{D_i} \sum_{\bar{i} \in I} A_{i\bar{i}} = \sum_{\alpha} \frac{\delta_{\Delta}(E_i - E_{\alpha})}{D_i} A_{\alpha\alpha}. \quad (\text{C2})$$

When $\delta_{\Delta}()$ is peaked around E_i , one obtains the first line of Eq. (23). This expression reproduces the result of global rotational invariance, but taking into account the energy dependence of the matrix elements; for a flat $\delta_{\Delta}()$ one recovers, indeed, the normalized trace of the entire matrix.

2. $n = 2$

We can distinguish two cases in the expression

$$\overline{A_{ij}^U A_{ji}^U} = \sum_{\substack{\bar{i}, \bar{j} \in I \\ \bar{k}, \bar{l} \in J}} \overline{U_{ii}^{(I)*} U_{kj}^{(J)*} U_{jj}^{(J)} U_{li}^{(I)}} A_{i\bar{j}} A_{\bar{k}\bar{l}}. \quad (\text{C3})$$

When $I = J$, we can use Weingarten calculus for dimension D_i . Expanding $\text{Wg}_{\alpha,\beta}(D_i)$ at the leading order in the dimension, one obtains:

$$\delta_{ij} \left(\frac{1}{D_i} \sum_{\bar{i} \in I} A_{i\bar{i}} \right)^2 + \frac{\delta_{I,J}}{D_i} \left(\frac{1}{D_i} \sum_{\bar{i}, \bar{k} \in I} A_{i\bar{k}} A_{\bar{k}\bar{i}} - \left(\frac{1}{D_i} \sum_{\bar{i} \in I} A_{i\bar{i}} \right)^2 \right). \quad (\text{C4})$$

When $I \neq J$, the average in Eq. (C3) factorizes in two

$$\overline{U_{ii}^{(I)*} U_{li}^{(I)}} \overline{U_{kj}^{(J)*} U_{jj}^{(J)}}.$$

Again, one can use Eq. (12) separately for the two averages, one with D_i and the other with D_j :

$$\frac{1 - \delta_{I,J}}{D_i D_j} \sum_{\substack{\bar{i} \in I \\ \bar{k} \in J}} A_{i\bar{k}} A_{\bar{k}\bar{i}}.$$

Summing these two cases together:

$$\begin{aligned} \overline{A_{ij}^U A_{ji}^U} &\simeq \delta_{ij} [K_1(E_i)]^2 + \frac{1}{D_i D_j} \sum_{\substack{\bar{i} \in I \\ \bar{k} \in J}} A_{i\bar{k}} A_{\bar{k}\bar{i}} \\ &- \left(\frac{1}{D_i} \sum_{\bar{i} \in I} A_{i\bar{i}} \right)^2 \frac{\delta_{I,J}}{D_i} \equiv \delta_{ij} [K_1(E_i)]^2 + K_2(E_i, E_j). \end{aligned} \quad (\text{C5})$$

Again, one can use equivalently smooth delta functions, so that

$$\begin{aligned} K_2(E_i, E_j) &= \sum_{\alpha\beta} A_{\alpha\beta} \frac{\delta_{\Delta}(E_i - E_{\alpha})}{D_i} A_{\beta\alpha} \frac{\delta_{\Delta}(E_j - E_{\beta})}{D_j} \\ &- [K_1(E_i)]^2 \sigma(E_i - E_j), \end{aligned} \quad (\text{C6})$$

where $\sigma(E_i - E_j)$ accounts for the smoothed version of $\delta_{I,J}/D_i$ on the mesoscopic energies. This expression shows that the simple loop is given by the second line of Eq. (23).

For the cactus $\overline{A_{ii}^U A_{ii}^U}$ we recover, from Eq. (C5), the factorization and the first subleading term with the right energy dependencies, as in Eq. (27a).

3. $n = 3$

For the expression

$$\overline{A_{ij}^U A_{jk}^U A_{ki}^U} = \sum_{\substack{\bar{i}, \bar{n} \in I \\ \bar{j}, \bar{k} \in J \\ \bar{l}, \bar{m} \in K}} \overline{U_{ii}^{(I)*} U_{kj}^{(J)*} U_{mn}^{(K)*} U_{jj}^{(J)} U_{lk}^{(K)} U_{ni}^{(I)}} A_{i\bar{j}} A_{\bar{k}\bar{l}} A_{\bar{m}\bar{n}}, \quad (\text{C7})$$

one has now to distinguish multiple cases, each computable with Eq. (12) choosing the right dimension, and expanding the Weingarten function.

We begin with the case of simple loop $i \neq j \neq k$. Equation (C7) needs to be considered in three different cases:

$I = J = K$ the Weingarten function has to be evaluated only for $\alpha = (123)$, since it is the only one respecting $i \neq j \neq k$, and expanded at the first order in $1/D_i$. The result is

$$\frac{\delta_{I,J}\delta_{I,K}}{D_i^2} \left(\frac{1}{D_i} \sum_{\bar{i}, \bar{k}, \bar{m} \in I} A_{\bar{i}\bar{k}} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{i}} - \frac{3}{D_i^2} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k}, \bar{m} \in I} A_{\bar{k}, \bar{m} \in I} A_{\bar{m}\bar{k}} + 2 \frac{(\sum_{\bar{i} \in I} A_{\bar{i}\bar{i}})^3}{D_i^3} \right). \quad (C8)$$

$I = J \neq K + 2$ cycl the average factorizes in two parts; one is easily computable directly with $\text{Wg}_{e,e}(D_k)$, and the other has to be computed only for $\alpha = (12)$. The result reads:

$$\frac{\delta_{I,J}(1 - \delta_{J,K})(1 - \delta_{K,I})}{D_i} \left(\frac{1}{D_i D_k} \sum_{\bar{i}, \bar{k} \in I, \bar{m} \in K} A_{\bar{i}\bar{k}} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{i}} - \frac{1}{D_i^2 D_k} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k} \in I, \bar{m} \in K} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}} \right). \quad (C9)$$

$I \neq J \neq K$ the average factorizes in three parts, each of them computable with $\text{Wg}_{e,e}$ for dimension D_i, D_j, D_k , respectively. The result is given by :

$$(1 - \delta_{I,J})(1 - \delta_{J,K})(1 - \delta_{K,I}) \frac{1}{D_i D_j D_k} \sum_{\bar{i} \in I, \bar{k} \in J, \bar{m} \in K} A_{\bar{i}\bar{k}} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{i}}. \quad (C10)$$

The different cases have to be summed together to get the final result:

$$\overline{A_{ij}^U A_{jk}^U A_{ki}^U} \simeq K_3(E_i, E_j, E_k) = \sum_{\alpha \in I, \beta \in J, \gamma \in K} A_{\alpha\beta} A_{\beta\gamma} A_{\gamma\alpha} - [K_1(E_i)] \frac{\delta_{I,J}}{D_i} \sum_{\alpha \in I, \beta \in K} A_{\alpha\beta} A_{\beta\alpha} + 2\text{cycl} + 2[K_1(E_i)]^3 \frac{\delta_{I,J}\delta_{J,K}}{D_i^2}.$$

The smoothing of this expression, done in the same way as explained above, gives the third line of Eq. (23). We move to the cactus $\overline{A_{ij}^U A_{ji}^U A_{ii}^U}$; Eq. (C7) becomes

$$\overline{A_{ij}^U A_{ji}^U A_{ii}^U} = \sum_{\substack{\bar{i}, \bar{j}, \bar{l}, \bar{m} \in I \\ \bar{j}, \bar{k} \in J}} \overline{U_{\bar{i}\bar{i}}^{(I)*} U_{\bar{k}\bar{j}}^{(J)*} U_{\bar{m}\bar{i}}^{(I)*} U_{\bar{j}\bar{j}}^{(J)} U_{\bar{l}\bar{i}}^{(I)} U_{\bar{n}\bar{i}}^{(I)} A_{\bar{i}\bar{j}} A_{\bar{k}\bar{l}} A_{\bar{m}\bar{n}}}, \quad (C11)$$

and needs to be considered in two different cases:

$I = J$ the Weingarten function has to be evaluated for $\alpha = (12)(3)$ and (123) , and expanded at the first order in $1/D_i$; not only, for each α one has to consider only the permutations $\beta < \alpha$ in order to have the first subleading contribution. We highlight that the other α s are not respecting the condition on the initial set of indices $\mathbf{i}$. It is remarkable to notice that the permutations to be considered are the ones on the geodesic between e and γ_3 , equal or bigger than $(12)(3)$, exactly as in the case of global rotational invariance. For instance, the permutation $(132) \notin [e, \gamma_3]$ and, indeed, does not respect the repetitions in the set of indices $\mathbf{i}$, even if with one cycle more than $(12)(3)$. The computation leads to

$$\frac{\delta_{I,J}}{D_i} \left(\frac{1}{D_i^2} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k}, \bar{m} \in I} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}} - \frac{(\sum_{\bar{i} \in I} A_{\bar{i}\bar{i}})^3}{D_i^3} + \frac{1}{D_i^2} \sum_{\bar{i}, \bar{k}, \bar{m} \in I} A_{\bar{i}\bar{k}} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{i}} + 2 \frac{(\sum_{\bar{i} \in I} A_{\bar{i}\bar{i}})^3}{D_i^4} - \frac{3}{D_i^3} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k}, \bar{m} \in I} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}} \right). \quad (C12)$$

$I \neq J$ the average factorizes in two parts; one easily computable directly with $\text{Wg}_{e,e}(D_j)$, and the other to be evaluated for both $\alpha = (12)$ and $\alpha = e$, always considering only $\beta < \alpha$. The result is

$$(1 - \delta_{I,J}) \left(\frac{1}{D_j D_i^2} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k} \in J, \bar{m} \in I} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}} + \frac{1}{D_j D_i^2} \sum_{\bar{i}, \bar{k} \in I, \bar{m} \in J} A_{\bar{i}\bar{k}} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{i}} - \frac{1}{D_j D_i^3} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k} \in J, \bar{m} \in I} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}} \right). \quad (C13)$$

Summing all together, one gets:

$$\overline{A_{ij}^U A_{ji}^U A_{ii}^U} \simeq \frac{1}{D_j D_i^2} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k} \in J, \bar{m} \in I} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}} - \frac{\delta_{I,J}}{D_i} \frac{(\sum_{\bar{i} \in I} A_{\bar{i}\bar{i}})^3}{D_i^3} \quad (\text{C14})$$

$$+ \frac{1}{D_j D_i^2} \sum_{\bar{i}, \bar{m} \in I, \bar{k} \in J} A_{\bar{i}\bar{k}} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{i}} + 2 \frac{\delta_{I,J}}{D_i} \frac{(\sum_{\bar{i} \in I} A_{\bar{i}\bar{i}})^3}{D_i^4} - \frac{1}{D_j D_i^3} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k} \in J, \bar{m} \in I} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}} - 2 \frac{\delta_{I,J}}{D_i} \frac{1}{D_i^3} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k}, \bar{m} \in I} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}} \\ \equiv K_2(E_i, E_j) K_1(E_i) + K_3(E_i, E_j, E_i), \quad (\text{C15})$$

which reproduces the right factorization and the first subleading contribution given by K_3 evaluated at equal mesoscopic energies, for coincident initial indices.

The case of the cactus $\overline{A_{ii}^U A_{ii}^U A_{ii}^U}$ is the easiest, but significant. There is only one average in Eq. (C7) to be considered, for which Weingarten calculus for dimension D_i can be employed, exactly as in the case of global rotational invariance. The Weingarten function has to be evaluated for all α 's, since all of them respect the condition on the initial set of indices $\mathbf{i}$, and have more cycles than e , but two considerations have to be made: firstly, (123) and (132) would give the second subleading correction, so they have to be excluded, for being coherent with the approximation at the leading order in the dimension made on the Weingarten function. Interestingly, (132) is respecting the repetitions of the indices in $\mathbf{i}$, even if it is not on the geodesic $[e, \gamma_3]$; indeed, as said in Appendix B, only for the first subleading correction it is sufficient to restrict to non-crossing partitions, but for higher orders this is not the case. The final result, in this case, is:

$$\overline{A_{ii}^U A_{ii}^U A_{ii}^U} \simeq \frac{(\sum_{\bar{i} \in I} A_{\bar{i}\bar{i}})^3}{D_i^3} + \frac{3}{D_i^3} \sum_{\bar{i} \in I} A_{\bar{i}\bar{i}} \sum_{\bar{k}, \bar{m} \in I} A_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}} \\ - 3 \frac{(\sum_{\bar{i} \in I} A_{\bar{i}\bar{i}})^3}{D_i^4} \equiv [K_1(E_i)]^3 + 3K_1(E_i)K_2(E_i, E_i).$$

Again, this reproduces the right factorization and the first subleading contribution.

4. $n = 4$

The same logic as outlined for the previous cases applies here, but the explicit derivation is much longer and therefore omitted. It is important to separate the different cases of macroscopic energies and, within each case, to factorize the averages before applying the Weingarten formalism. Only the relevant α 's should be considered, the function should be approximated to leading order, and only terms with $\beta < \alpha$ retained.

APPENDIX D: LINK BETWEEN LOCAL FREE CUMULANTS AND OPERATOR-VALUED FREE CUMULANTS

We now show how Eqs. (25) emerge within the framework of operator-valued free probability. To this end, let us reconsider the example introduced in Appendix A 2, namely a four-block matrix and the choice of $\mathcal{D}$ as the subalgebra of block-diagonal matrices generated by the projectors $\hat{\Pi}_{I=1,2}$ onto the individual blocks, see Eq. (A3). Generalizing this construction to an arbitrary number of blocks provides the natural setting in which operator-valued free cumulants, valued by these projectors, can be related to local free cumulants. Indeed, local free cumulants arise when a physical observable A in the energy basis is regarded as a $D \times D$ matrix partitioned into blocks according to the energies of the states, with each block being independently reshuffled by a local rotation. However, unlike operator-valued free cumulants, which are matrix-valued, local free cumulants are scalar quantities. In fact, we now show explicitly that they correspond to the diagonal entries of the blocks of the operator-valued free cumulants. Furthermore, a suitable rescaling of the projectors is required to recover Eqs. (25).

We define the normalized projectors on the two diagonal blocks as $\hat{\sigma}_I := \hat{\Pi}_I / D_I$ being D_I the dimension of $\mathcal{A}_{II}$.

For the first-order free cumulant, we have [Eq. (A4)]:

$$\langle A \rangle_{\mathcal{D}} = \kappa_1^{\mathcal{D}}(A) = \text{diag}(\langle \mathcal{A}_{11} \rangle \mathbb{I}, \langle \mathcal{A}_{22} \rangle \mathbb{I}). \quad (\text{D1})$$

The value on the diagonal of each block can be expressed as the global trace of the matrix projected onto the corresponding diagonal block via the projector:

$$\langle \mathcal{A}_{II} \rangle = \text{Tr}(A \hat{\sigma}_I). \quad (\text{D2})$$

Indeed, this corresponds simply to the normalized trace of the diagonal block of the original matrix. We call the r.h.s. of Eq. (D2) as *generalized moment*, and we introduce the first *generalized free cumulant* as

$$K_1(\mathcal{A}_{II}) = \text{Tr}(A\hat{\sigma}_I). \quad (\text{D3})$$

These generalized quantities are scalars and explicitly retain the dependence on the specific diagonal block.

We can proceed analogously for the second-order case, for which we report Eqs. (A5), (A7):

$$\begin{aligned} \langle A\hat{\sigma}_1 A \rangle_{\mathcal{D}} &= \text{diag}(\langle \mathcal{A}_{11}^2 \rangle / D_1 \mathbb{I}, \langle \mathcal{A}_{21} \mathcal{A}_{12} \rangle / D_1 \mathbb{I}), \\ \langle A\hat{\sigma}_2 A \rangle_{\mathcal{D}} &= \text{diag}(\langle \mathcal{A}_{12} \mathcal{A}_{21} \rangle / D_2 \mathbb{I}, \langle \mathcal{A}_{22}^2 \rangle / D_2 \mathbb{I}), \\ \kappa_2^{\mathcal{D}}(A\hat{\sigma}_1, A) &= \text{diag}((\langle \mathcal{A}_{11}^2 \rangle - \langle \mathcal{A}_{11} \rangle^2) / D_1 \mathbb{I}, \langle \mathcal{A}_{21} \mathcal{A}_{12} \rangle / D_1 \mathbb{I}), \\ \kappa_2^{\mathcal{D}}(A\hat{\sigma}_2, A) &= \text{diag}(\langle \mathcal{A}_{12} \mathcal{A}_{21} \rangle / D_2 \mathbb{I}, (\langle \mathcal{A}_{22}^2 \rangle - \langle \mathcal{A}_{22} \rangle^2) / D_2 \mathbb{I}). \end{aligned}$$

We now define the second-order generalized moments as

$$\begin{aligned} \langle \mathcal{A}_{II}^2 \rangle / D_I &= \text{Tr}(A\hat{\sigma}_I A\hat{\sigma}_I) \\ \langle \mathcal{A}_{IJ} \mathcal{A}_{JI} \rangle / D_J &= \text{Tr}(A\hat{\sigma}_I A\hat{\sigma}_J), \end{aligned} \quad (\text{D4})$$

which are again the diagonal values of the operator-valued free cumulants. The generalized second-order free cumulants follow as

$$\begin{aligned} K_2(\mathcal{A}_{II}, \mathcal{A}_{II}) &= \text{Tr}(A\hat{\sigma}_I A\hat{\sigma}_I) - \text{Tr}(A\hat{\sigma}_I)^2 \\ &\equiv \frac{1}{D_I} (\langle \mathcal{A}_{II}^2 \rangle - \langle \mathcal{A}_{II} \rangle^2), \\ K_2(\mathcal{A}_{II}, \mathcal{A}_{JJ}) &= \text{Tr}(A\hat{\sigma}_I A\hat{\sigma}_J) \\ &\equiv \frac{1}{D_J} \langle \mathcal{A}_{IJ} \mathcal{A}_{JI} \rangle, \end{aligned} \quad (\text{D5})$$

which compactly is

$$K_2(\mathcal{A}_{II}, \mathcal{A}_{JJ}) = \text{Tr}(A\hat{\sigma}_I A\hat{\sigma}_J) - \frac{\delta_{I,J}}{D_I} \text{Tr}(A\hat{\sigma}_I)^2. \quad (\text{D6})$$

One should note that these formulas extend straightforwardly to a generic number of blocks in the initial matrix. The resulting free cumulants are scalars, but they explicitly encode correlations between different diagonal blocks. Moreover, an additional Kronecker δ structure is inherited from operator-valued free probability.

In the case of an observable A expressed in the energy basis, the generalized free cumulants capture correlations between different diagonal blocks, each associated with a given mesoscopic energy. In this setting, the projector $\hat{\sigma}_I = \sum_{\alpha} |\alpha\rangle \langle \alpha| \frac{\delta_{\Delta}(E_i - E_{\alpha})}{D_I}$ selects the set of states with energy E_i , where i labels a representative state within the block. We may therefore replace the block index I with the corresponding energy E_i , and identify $\delta_{I,J}/D_I$ with $\sigma(E_i - E_j) = \frac{\delta_{\Delta}(E_i - E_j)}{D_I}$. In this representation, the first two cumulants read

$$K_1(E_i) = \text{Tr}(\hat{\sigma}_{E_i} A)$$

$$K_2(E_i, E_j) = \text{Tr}(\hat{\sigma}_{E_i} A \hat{\sigma}_{E_j} A) - [\text{Tr}(\hat{\sigma}_{E_i} A)]^2 \sigma(E_i - E_j). \quad (\text{D7})$$

These expressions correspond to Eqs. (23a)–(23b). Upon computing higher orders, one obtains the *moments-cumulants formula* in Eq. (25). This establishes the identification of the local free cumulants with the generalized free cumulants here introduced, namely the diagonal-block components of the operator-valued free cumulants.

The equivalence has been shown explicitly for the toy model with blocks. To obtain smooth functions of the energy, one must smooth each block around its associated energy: the factor $\delta_{\Delta}(E_i - E_{\alpha})$ is then understood as a function peaked around E_i , rather than a sharp box. In this regime, the connection to operator-valued free probability becomes more subtle. In particular, one must appropriately integrate the local cumulants over energy in order to recover operator-valued cumulants. This was carried out in [60], even if in a different setting.

APPENDIX E: A DIFFERENT APPROACH FOR GLOBAL ROTATIONAL INVARIANCE: LOW RANK HCIZ INTEGRAL

The Harish-Chandra-Itzykson-Zuber (HCIZ) integral [118,119] $\mathcal{I}_{\beta_{\text{RMT}}}(A, B)$ is defined as

$$\mathcal{I}_{\beta_{\text{RMT}}}(A, B) = \int_{G(D)} \mathcal{D}\Omega e^{\frac{\beta_{\text{RMT}} D}{2} \text{Tr} B \Omega A \Omega^{\dagger}}, \quad (\text{E1})$$

where the integral is over the Haar measure of the compact group $\Omega \in G(D) = O(D)$ (orthogonal) or $U(D)$ (unitary) in D dimensions and A, B are arbitrary $D \times D$ symmetric or hermitian matrices [120]. The parameter β_{RMT} is the usual Dyson “inverse temperature,” with $\beta_{\text{RMT}} = 1$ or 2, respectively for the two groups.

Several results are known for this integral but in the following we will be interested in a simple yet beautiful explicit prediction which is available when one of the matrices has lower rank $n \ll D$. Precisely, let us assume that B has n eigenvalues $\lambda_1^B, \dots, \lambda_n^B$ and $D - n$ zero eigenvalues. Then we have [121,122]:

$$\mathcal{I}_{\beta_{\text{RMT}}}(A, B) = \exp \left[\frac{D\beta_{\text{RMT}}}{2} \sum_{i=1}^n \mathcal{W}_A(\lambda_i^B) \right] \quad (\text{E2})$$

where $\mathcal{W}_A$ is the primitive of the $\mathcal{R}$ -transform of A :

$$\mathcal{W}_A(z) = \int_0^z dx \mathcal{R}_A(x). \quad (\text{E3})$$

with $\mathcal{R}_A(z)$ defined as the generating function of free cumulants:

$$\mathcal{R}_A(z) = \sum_{k=0}^{\infty} \kappa_{k+1}^A z^k. \quad (\text{E4})$$

We note that $\sum_{i=1}^n \mathcal{W}_A(\lambda_i^B)$ can also be written as $\text{Tr } \mathcal{W}_A(B)$ and the exponent in Eq. (E2) can be written as

$$\mathcal{W}_A(B) = \sum_{k=1}^{\infty} \frac{1}{k} \kappa_k^A \text{Tr}(B^k). \quad (\text{E5})$$

For an intuitive derivation of Eq. (E2) we refer to [120].

We now sketch how the predictions for the leading terms derived in the main section in the unitary ensemble can be generalized also to the orthogonal case. These results and the proof that we present are discussed in [86]. Generalizing the previous notation, let us now consider a matrix $A^O = O^\dagger A O$ with O a Haar distributed orthogonal matrix (but we could consider different universality classes), and A a matrix with limiting spectral function ρ^A . The result (E2) can be used to show that for any $n \geq 1$ and any set of pairwise distinct indices $i_1, \dots, i_n$, one has:

$$\lim_{D \rightarrow \infty} [L_n^{(D)} \equiv D^{n-1} \overline{A_{i_1 i_2}^O A_{i_2 i_3}^O \cdots A_{i_{n-1} i_n}^O A_{i_n i_1}^O}] = \kappa_n^A \quad (\text{E6})$$

where the average is over the orthogonal matrix. Note that the free cumulants of A are the same as those of A^O as the two matrices share the same spectral properties.

In order to simplify the following calculation, we assume that $(i_1, \dots, i_n) = (1, \dots, n)$, since the desired result does not depend on the particular choice of indices (as is clear by rotational invariance).

We note that we can use (E1), and in particular its low rank limit, as a generating function for the matrix elements of A^O . The cases $n = 1$ and $n = 2$ have to be treated separately, but one can generalize the case we consider here for $n \geq 3$ where we can write $L_n^{(D)}$ as

$$L_n^{(D)} = \frac{1}{D} \prod_{l=1}^n \frac{\partial}{\partial b_l} \left[\int_{O(N)} \mathcal{D}O e^{\frac{D}{2} \text{Tr}[B(\mathbf{b}) O A O^\dagger]} \right]_{\mathbf{b}=0}, \quad (\text{E7})$$

in which we denoted $\mathbf{b} \equiv (b_1, \dots, b_n)$ and the following symmetric block matrix of rank n :

$$B(\mathbf{b}) \equiv \begin{pmatrix} B_1(\mathbf{b}) & (0) \\ (0) & (0) \end{pmatrix}, \quad (\text{E8})$$

in which $B_1(\mathbf{b})$ is

$$B_1(\mathbf{b}) \equiv \begin{pmatrix} 0 & b_1 & 0 & \cdots & 0 & b_n \\ b_1 & 0 & b_2 & \cdots & 0 & 0 \\ 0 & b_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & b_{n-1} \\ b_n & 0 & 0 & \cdots & b_{n-1} & 0 \end{pmatrix}. \quad (\text{E9})$$

Upon taking derivatives in (E7), using the low rank result (E5), and the form (E9), one can recover Eq. (E6).

With the same strategy one can construct a $B(\mathbf{b})$ matrix which is used to compute averages of cactus diagrams. As deriving the exponent in (E5) allows to gain a factor D at the leading order it is clearly convenient to decompose the cactus in as many leaves (cycles) as possible, noting that this decomposition is unique, and then the argument follows the same lines as for simple loops.

The same approach can be generalized to the unitary case (see Ref. [86]) and the leading terms of the expectations turns out to be the same for real and complex matrices.

It turns out that the low rank integral captures also the corrections with scaling $1/D$ smaller than the leading term for the cactus diagrams. At this order, however, one might expect some differences in the two universality classes. In fact, different factors appear in front of the expressions of the subleading corrections, beyond the one provided by the free cumulants, for products involving diagonal matrix elements in the case of orthogonal matrices. One example is the cactus diagram which describes the fluctuations of diagonal matrix elements:

$$\overline{A_{ii}^\Omega A_{ii}^\Omega} \simeq \overline{A_{ii}^\Omega}^2 + \frac{2}{\beta_{\text{RMT}}} \overline{A_{ij}^\Omega A_{ji}^\Omega}, \quad (\text{E10})$$

for $\Omega = O$ or $\Omega = U$, as it was noticed for instance in [48,89].

APPENDIX F: ADDITIONAL NUMERICAL RESULTS

In the first section of this Appendix, we report some additional details on the implementations of the numerics; we remind the reader that all the codes and data can be found in [95]. Successively, we present numerical results for slightly different Hamiltonians than the one in the main text: we still consider a non-integrable periodically driven spin chain, described by Hamiltonian in (32), and by the Floquet operator in (34), but we slightly change the form of H_0 or V_0 in (33). Firstly, we break translational invariance, in order to check our assumptions for different observables, then, we restore the time reversal symmetry, so that the symmetry class of the Hamiltonian is orthogonal instead of unitary.

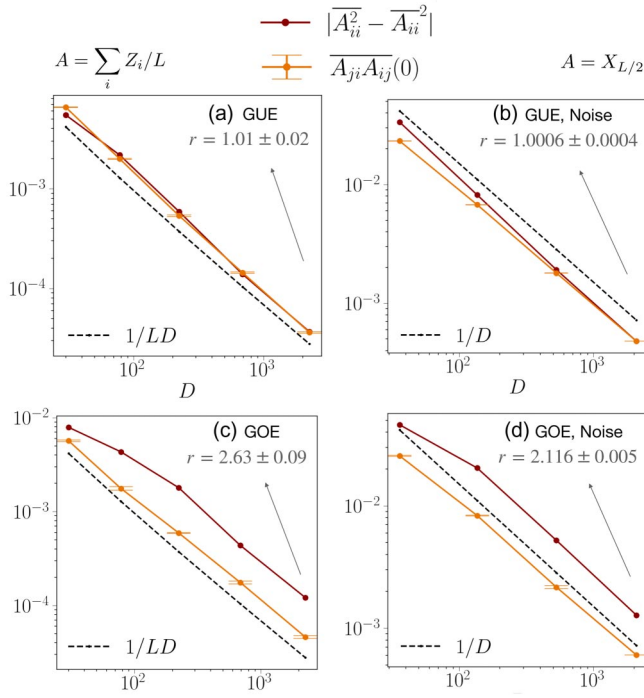


FIG. 5. Numerical comparison between the Floquet systems with and without time reversal symmetry, that is, belonging to the orthogonal or the unitary symmetry class, with and without disorder, presented in Appendix F 2. We numerically reproduce equation Eq. (E10). The error bar is given by averaging $\overline{A_{ij}A_{ji}(0)}$ over different values of the smoothing parameter Δ , as explained in the text. For the first column $L \in [8, 10, 12, 14, 16]$, while for the second $L \in [6, 8, 10, 12]$. We compare $|\overline{A_{ii}^2} - \overline{A_{ii}}|^2$ (brown curve) with the respective subleading correction $|\overline{A_{ij}A_{ji}(0)}|$ (orange curve), as functions of D , and we calculate the ratio $r = |\overline{A_{ii}^2} - \overline{A_{ii}}|^2 / |\overline{A_{ij}A_{ji}(0)}|$ for the biggest system size for each figure. All the curves well reproduce the expected scaling of $1/LD$ for the system without disorder, and $1/D$ for the system with disorder. In (a) and (b) we notice that the two curves almost coincide, reproducing the result for the unitary symmetry class (27a). In (c) and (d), instead, $|\overline{A_{ii}^2} - \overline{A_{ii}}|^2$ is approximately twice as large as the subleading one, as expected for the orthogonal symmetry class (F1).

1. Different observables

One can break the translational invariance, considering, for instance, a different amplitude for the longitudinal magnetic field on each site in Eq. (33):

$$H_z^{\text{disorder}} = J \sum_i Z_i Z_{i+1} + \sum_i h_i Z_i.$$

Each h_i is independently sampled from a uniform distribution centered at $\bar{h} = 0.905$ and spanning the interval $[\bar{h} - \delta_h, \bar{h} + \delta_h]$, with width $\delta_h = 0.1\sqrt{3}$. In this case, we can consider as observable the magnetization on a single site; we choose $A = X_{L/2}$. We perform the same

analysis as in Sec. VIC of the main, evaluating the averages of the products of matrix elements involved in Eqs. (36a)–(36b), through the expressions (F2).

All results are plotted in Fig. 6. For a fixed number of spin $L = 12$, we compare the right-hand side (purple curve) of Eqs. (36a)–(36b) with the respective leading order (pink curve), and then we subtract the leading to the right-hand side, in order to compare this difference (brown curve) with the prediction for the subleading (orange curve). Moreover, we study the scaling with the subspace dimension D for each of these curves, for a fixed $\omega^* \approx -2.13$. Since $\langle X_{L/2} \rangle = 0$, the leading order of $\overline{A_{ij}A_{ji}A_{ii}A_{ii}}(\omega)$ is zero, therefore the average follows directly the subleading contribution $\overline{A_{ij}A_{ji}}(\omega)\overline{A_{il}A_{li}}(0)$, and scales as $1/D^2$, as expected. On the contrary, $\overline{A_{ij}A_{ji}A_{il}A_{li}}(\omega, \omega)$ is very well factorized in $\overline{A_{ij}A_{ji}}(\omega)\overline{A_{il}A_{li}}(\omega)$, scaling as $1/D^2$, and the subleading well follows the expected curve $\overline{A_{ij}A_{jk}A_{kl}A_{li}}(\omega, -\omega, \omega)$, scaling as $1/D^3$.

2. Orthogonal symmetry class

One can restore the time reversal symmetry considering no H_M interactions, i.e. $d = 0$ in Eq. (33). In this case, the system belongs to the orthogonal symmetry class, instead of the unitary [98]. As mentioned in Appendix E, it is well known that the predictions of standard ETH for the subleading of the cacti for $n = 2$ depend on the symmetry class [7,89], see Eq. (E10). For the orthogonal:

$$\overline{A_{ii}A_{ii}} - \overline{A_{ii}}^2 \simeq 2\overline{A_{ij}A_{ji}}(0). \quad (\text{F1})$$

Therefore, we expect to have discrepancies with respect to the unitary case for the subleading of higher n , in particular when diagonal matrix elements are involved. The origin of the differences is in the rotational invariance: the observable A would be invariant under local orthogonal transformations, as a consequence of the pseudorandomness of real eigenvectors of the Hamiltonian. Therefore, the computation of averages of products of matrix elements could not be carried out via the Weingarten calculus with permutations [Eq. (12)], as we did instead in the unitary case. Providing an analytical and general prediction of how the subleading corrections would change for the orthogonal case goes beyond the scope of this paper, but numerically we can perform again the same analysis as in Sec. VIC of the main in order to observe the discrepancies. Firstly, in Fig. 5 for both the model given by the (33) and the variation with disorder presented in Appendix F 1, we study the factorization in Eq. (F1). Since we are considering a Floquet model, the products of matrix elements on the left-hand side are constant for a fixed L , therefore the averages are computed as normal means over all the diagonal elements and plotted as a function of the dimension D . We observe the difference

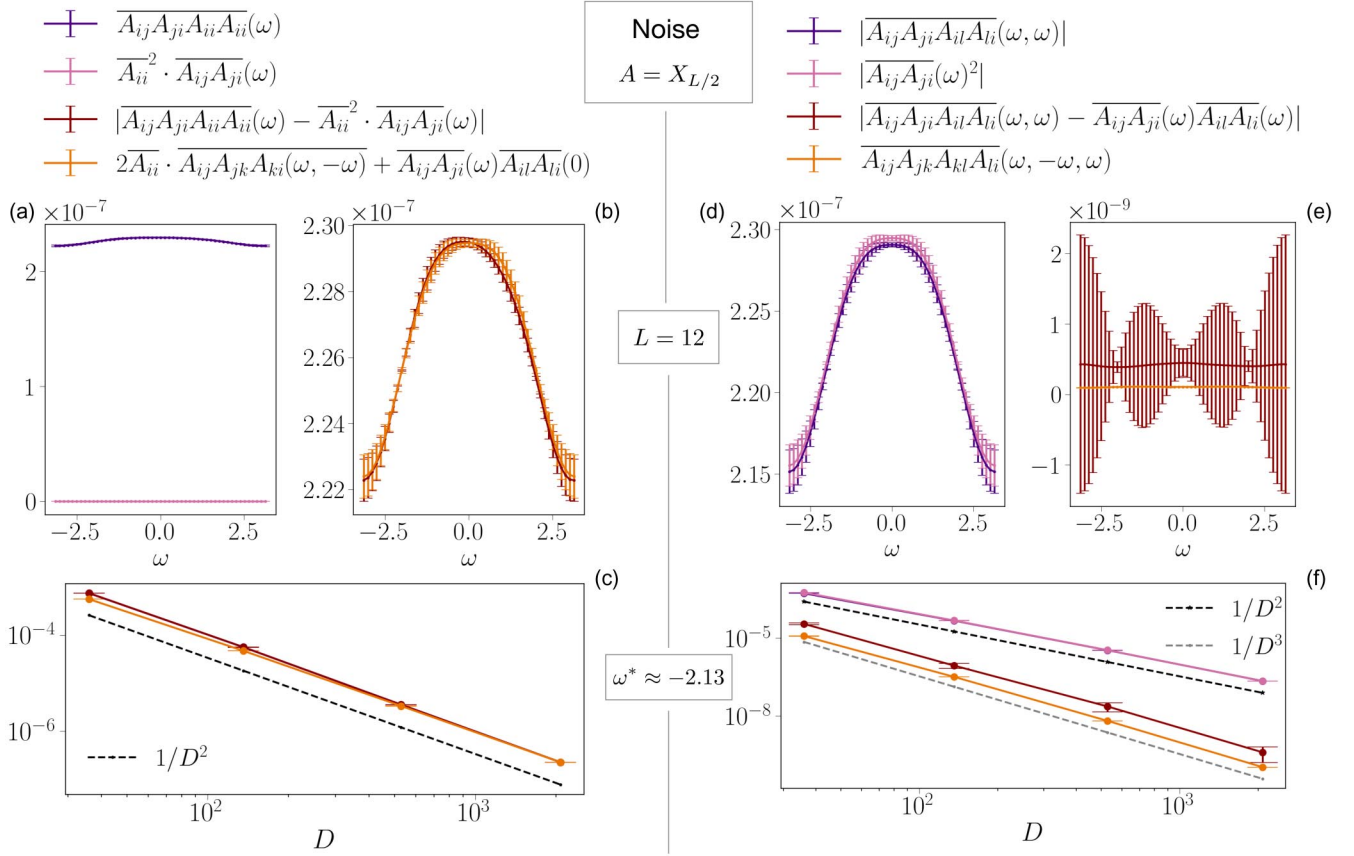


FIG. 6. Numerical results for the Floquet system with disorder presented in Appendix F 1, and for the observable $A = X_{L/2}$. Similarly to Fig. 4, we numerically reproduce equations Eqs. (36a)–(36b), for $\omega_1 = \omega_2 = \omega$ chosen in $[-\pi, \pi]$. The error bar is given by averaging each product, for each ω , over different values of the smoothing parameter Δ , as explained in the text. (a),(b),(c) refer to (36a), while (d),(e),(f) to (36b). In (a),(d) we compare the right-hand side (purple curve) with the respective leading order (pink curve) for fixed $L = 12$; for (a) the leading order is zero, while in (d) the data very well reproduce the expected leading order factorization. In (b),(e) we compare the difference between the right-hand side and the leading contribution (brown curve) with the respective subleading correction (orange curve) for fixed $L = 12$. The data verify the subleading correction to the factorization. In (c),(f) we represent the scalings of all the curves as a function of D . The scalings were studied for values $L \in [6, 8, 10, 12]$, and for a fixed value of $\omega^* \approx -2.13$. While all the curves in (c) scale as $1/D^2$, since the leading order is zero, in (f) we observe the different scalings between the leading (purple and pink curves), as $1/D^2$ and the subleading (brown and orange curves), as $1/D^3$.

between the unitary ($d = 1$) and the orthogonal case ($d = 0$) in the proportionality factor $r = |\overline{A_{ii}A_{ii}} - \overline{A_{ii}^2}|/|\overline{A_{ij}A_{ji}}(0)|$ between the left-hand side (brown curve) and the right-hand side (orange curve), that is respectively close to one and close to two. The proportionality factors are explicitly reported in Fig. 5, for the biggest system size of each case.

We then perform the same analysis as in Sec. VIC and Appendix F 1, plotting the result in Fig. 7 for $d = 0$. Specifically, we compare the right-hand side (purple curve) of Eqs. (36a)–(36b) with the respective leading order (pink curve), and then we subtract the leading to the right-hand side, in order to compare this difference (brown curve) with the prediction for the subleading (orange curve). We notice that the factorizations are always very well reproduced, while for the subleading contributions,

the same difference observed in Fig. 7 persists while considering the case of Eq. (36a). The proportionality factors are explicitly reported in Fig. 7 and are obtained averaging over all the ω . This discrepancy is not present in the case of Eq. (36b) which, indeed, does not involve the diagonal.

3. Numerical implementation

Regarding more specifically the numerical implementation, we write the empirical averages used for the numerical calculation of Eqs. (36a)–(36b), in the case of a Floquet system.

$$\overline{A_{ij}A_{ji}A_{ii}A_{ii}}(\omega) = \frac{\sum_{i \neq j} A_{ij} \delta_{\omega}^{\text{per}}(\nu_i, \nu_j) A_{ji} A_{ii} A_{ii}}{\sum_{i,j} \delta_{\omega}^{\text{per}}(\nu_i, \nu_j)}, \quad (\text{F2a})$$

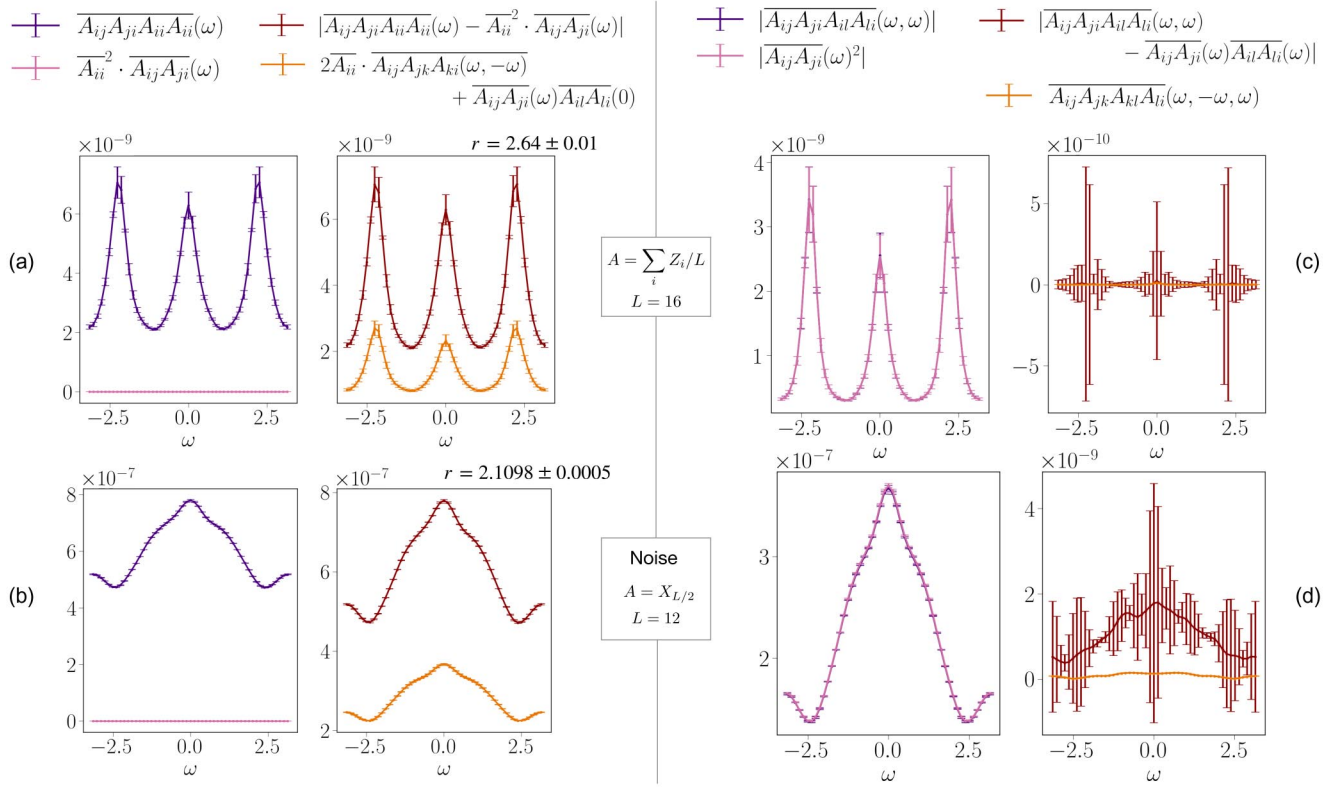


FIG. 7. Numerical results for the Floquet system belonging to the orthogonal symmetry class presented in Appendix F 2. In particular, (a) and (c) refer to the system in Eq. (33), while (b) and (d) to the variation of the system with disorder. We numerically reproduce equations Eqs. (36a)–(36b), for $\omega_1 = \omega_2 = \omega$ chosen in $[-\pi, \pi]$. The error bar is given by averaging each product, for each ω , over different values of the smoothing parameter Δ , as explained in the text. (a),(b) refer to (36a), while (c),(d) to (36b). We compare the right-hand side (purple curve) with the respective leading order (pink curve) for fixed system size; the data very well reproduce the expected leading order factorization. Then, we compare the difference between the right-hand side and the leading contribution (brown curve) with the respective subleading correction (orange curve) for fixed system size. In (a) and (b) we notice that the first curve is approximately twice as large as the subleading one, in contrast to the unitary case (see Figs. 4, 6). The ratio r between these two curves, averaged over ω , is respectively $r = 2.66 \pm 0.01$ and $r = 2.095 \pm 0.007$ for the two systems considered, in accordance with Fig. 5. In (c) and (d), instead, we notice no difference between the unitary and the orthogonal case.

$$\overline{A_{ij}A_{ji}A_{il}A_{li}}(\omega_1, \omega_2) = D \frac{\sum_{i \neq j \neq l} A_{ij} \delta_{\omega_1}^{\text{per}}(\nu_i, \nu_j) A_{ji} A_{il} \delta_{\omega_2}^{\text{per}}(\nu_i, \nu_l) A_{li}}{\sum_{i,j} \delta_{\omega_1}^{\text{per}}(\nu_i, \nu_j) \sum_{i,l} \delta_{\omega_2}^{\text{per}}(\nu_i, \nu_l)}, \quad (\text{F2b})$$

$$\overline{A_{ii}} = \frac{1}{D} \sum_i A_{ii}, \quad \overline{A_{ij}A_{ji}}(\omega) = \frac{\sum_{i \neq j} A_{ij} \delta_{\omega}^{\text{per}}(\nu_i, \nu_j) A_{ji}}{\sum_{i,j} \delta_{\omega}^{\text{per}}(\nu_i, \nu_j)}, \quad (\text{F2c})$$

$$\overline{A_{ij}A_{jk}A_{ki}}(\omega, -\omega) = D \frac{\sum_{i \neq j \neq k} A_{ij} \delta_{\omega}^{\text{per}}(\nu_i, \nu_j) A_{jk} \delta_{-\omega}^{\text{per}}(\nu_j, \nu_k) A_{ki}}{\sum_{i,j} \delta_{\omega}^{\text{per}}(\nu_i, \nu_j) \sum_{j,k} \delta_{-\omega}^{\text{per}}(\nu_j, \nu_k)}, \quad (\text{F2d})$$

$$\overline{A_{ij}A_{jk}A_{kl}A_{li}}(\omega_1, -\omega_1, \omega_2) = D^2 \frac{\sum_{i \neq j \neq k \neq l} A_{ij} \delta_{\omega_1}^{\text{per}}(\nu_i, \nu_j) A_{jk} \delta_{-\omega_1}^{\text{per}}(\nu_j, \nu_k) A_{kl} \delta_{\omega_2}^{\text{per}}(\nu_k, \nu_l) A_{li}}{\sum_{i,j} \delta_{\omega_1}^{\text{per}}(\nu_i, \nu_j) \sum_{j,k} \delta_{-\omega_1}^{\text{per}}(\nu_j, \nu_k) \sum_{k,l} \delta_{\omega_2}^{\text{per}}(\nu_k, \nu_l)}. \quad (\text{F2e})$$

Moreover, when we implemented the smoothing through the Gaussian function

$$\delta_{\Delta}(\omega - (\nu_{\alpha} - \nu_{\beta})) = \frac{1}{\sqrt{\pi}\Delta} \exp\left(-\frac{(\omega - (\nu_{\alpha} - \nu_{\beta}))^2}{\Delta}\right),$$

we calculated each matrix elements' moment for multiple values of Δ s. In particular, we fixed intervals of Δ , based on the model considered (unitary or orthogonal symmetry class) and on the number of spins L , as shown in the Table below. Inside each interval we chose 3 different values of Δ , we calculated each matrix elements' moment for each Δ , and we averaged them in order to have a final function, with the respective error bars.

Δ 's intervals	$L = 6, 8$	$L = 10$	$L = 12$	$L = 14, 16$
Unitary	[2, 4]	[0.5, 1.5]	[0.5, 1.5]	[0.1, 0.7]
Orthogonal	[0.3, 0.7]	[0.1, 0.5]	[0.07, 0.1]	[0.01, 0.07]

APPENDIX G: THERMAL FREE CUMULANTS AND SUBLEADING CORRECTIONS

In this Appendix we focus on thermal correlation functions [cf. Eq. (42)] of Floquet dynamics [123], for which $\langle \bullet \rangle_{\beta=0} = \text{Tr}(\bullet)/D$ indicates the thermal average. In the following we will drop the subscript $\beta = 0$. Moreover, we focus here explicitly on the case of the two-point correlation function $\langle A(t)A \rangle$ and of the four-point correlation function $\langle A(t)A(t)AA \rangle$; other correlations can be treated similarly. For the thermal moments one can define thermal free cumulants between observables at different times $\vec{t} = (t_1, t_2, \dots, t_n)$ through the moment-cumulant formula (see Eq. (17)) as:

$$\langle A(t_1)A(t_2)\dots A(t_n) \rangle = \sum_{\pi \in NC(n)} k_{\pi}(A(t_1), \dots, A(t_n)), \quad (\text{G1})$$

where k_{π} is a product of free cumulants over the blocks of partitions π , as in Eq. (17). We will denote the thermal free cumulants as $k_n(\vec{t})$.

These free cumulants are connected correlation functions, but in the ETH setting they were proven at the leading order to be given only by sums over simple loops [32]

$$k_n(\vec{t}) \approx k_n^{\text{ETH}}(\vec{t}) = \frac{1}{D} \sum_{i_1 \neq \dots \neq i_n} A_{i_1 i_2} \dots A_{i_n i_1} e^{i\vec{\omega}\vec{t}} \quad (\text{G2})$$

$$\simeq \frac{1}{D} \sum_{i_1 \neq \dots \neq i_n} K_n(\vec{\omega}) e^{i\vec{\omega}\vec{t}}. \quad (\text{G3})$$

In the second line, we have also written the expression in terms of the local free cumulants, which depend only on the energy differences $\vec{\omega} = (\omega_1, \dots, \omega_n)$ with $\omega_{\alpha} \equiv \omega_{i_{\alpha} i_{\alpha+1}} = E_{i_{\alpha}} - E_{i_{\alpha+1}}$. Eq. (4) is valid because, via ETH, crossing contributions are shown to be suppressed and non-crossing diagrams factorize at the leading order. The subleading corrections derived in this work allow us to

go beyond this leading-order result and, in particular, to better predict the long-time plateaus of thermal correlations and free cumulants. Such corrections can become relevant whenever leading-order contributions cancel. More generally, they can control the finite-size scaling of the residual long-time values of thermal free cumulants. Since the Hilbert-space dimension D grows exponentially with system size, these plateaus are themselves exponentially suppressed in the thermodynamic limit. This establishes a quantitative connection between finite-size corrections and the emergence of long-time freeness in chaotic quantum dynamics. We now show some examples.

The first-order thermal free cumulant is time independent and exactly equal to the ETH free cumulant, as well as to the local free cumulant:

$$k_1 \equiv \langle A \rangle = k_1^{\text{ETH}} = K_1. \quad (\text{G4})$$

For the two-time correlation function,

$$\begin{aligned} \langle A(t)A \rangle &= \frac{1}{D} \sum_{i,j} A_{ij} A_{ji} e^{i\omega_{ij}t} \\ &= \frac{1}{D} \sum_{i \neq j} A_{ij} A_{ji} e^{i\omega_{ij}t} + \frac{1}{D} \sum_i A_{ii} A_{ii} \\ &= \frac{1}{D} \sum_{i \neq j} K_2(\omega_{ij}) e^{i\omega_{ij}t} + K_1^2 + K_2(0) + \mathcal{O}(D^{-2}) \\ &= k_2^{\text{ETH}}(t, 0) + (k_1^{\text{ETH}})^2 + K_2(0) + \mathcal{O}(D^{-2}), \end{aligned} \quad (\text{G5})$$

and therefore for the second-order thermal free cumulant

$$k_2(t, 0) \equiv \langle A(t)A \rangle - \langle A \rangle^2 = k_2^{\text{ETH}}(t, 0) + K_2(0) + \mathcal{O}(D^{-2}). \quad (\text{G6})$$

ETH free cumulants are oscillatory contributions and therefore vanish upon time averaging under the non-resonance condition, as shown in Ref. [43]. The long-time plateaus are thus entirely determined by the non-oscillating contributions, giving

$$[\langle A(t)A \rangle]_{\infty} \simeq (k_1^{\text{ETH}})^2 + K_2(0), \quad (\text{G7})$$

$$[k_2(t, 0)]_{\infty} \simeq K_2(0). \quad (\text{G8})$$

Since $K_2 \propto D^{-1}$, the plateau of the second-order thermal free cumulant is a finite-size effect that scales as D^{-1} . Similar considerations apply to fluctuations around the long-time plateaus of thermal free cumulants. These fluctuations are of order $\mathcal{O}(D^{-2})$; see, e.g., Ref. [46].

We now turn to correlators of the form

$$C_4(t) \equiv \langle A(t)A(t)AA \rangle = \frac{1}{D} \sum_{i,j,k,l} A_{ij}A_{jk}A_{kl}A_{li} e^{i(E_i - E_k)t}. \quad (\text{G9})$$

Equivalently $C_4(t) = \langle A^2(t)A^2 \rangle$. Under the non-resonance condition, the infinite-time average selects the terms with $i = k$. We separate the surviving contributions according to the possible coincidences among the indices i, j , and l :

$$\begin{aligned} [C_4(t)]_\infty &= \frac{1}{D} \sum_i A_{ii}^4 + \frac{2}{D} \sum_{i \neq j} A_{ii}A_{ii}A_{ij}A_{ji} \\ &\quad + \frac{1}{D} \sum_{i \neq j \neq l} A_{ij}A_{ji}A_{il}A_{li} + \frac{1}{D} \sum_{i \neq j} A_{ij}A_{ji}A_{ij}A_{ji}, \end{aligned} \quad (\text{G10})$$

where the last term is the crossing contribution. For convenience, we take A to be traceless, namely $K_1 = 0$. In this case, the first term in Eq. (G10) is of order $\mathcal{O}(D^{-2})$ and can therefore be neglected at the order considered here, while the second term contributes first at subleading order. For the leading part of the third term, using the ETH factorization and accounting for the restriction $j \neq l$, we have

$$\begin{aligned} &\frac{1}{D} \sum_{i \neq j \neq l} K_2(\omega_{ij})K_2(\omega_{il}) \\ &= \frac{1}{D} \sum_{i \neq j \neq l} K_2(\omega_{ij})K_2(\omega_{il}) - \frac{1}{D} \sum_{i \neq j} K_2(\omega_{ij})K_2(-\omega_{ij}) \\ &\simeq [k_2^{\text{ETH}}(0,0)]^2 - \frac{1}{D} \sum_{i \neq j} K_2(\omega_{ij})K_2(-\omega_{ij}) \end{aligned} \quad (\text{G11})$$

The last term in Eq. (G10) is instead the crossing contribution and gives $\frac{2}{D} \sum_{i \neq j} K_2(\omega_{ij})K_2(-\omega_{ij})$, following from Eq. (B17).

Combining all contributions, we obtain

$$\begin{aligned} [C_4(t)]_\infty &\simeq [k_2^{\text{ETH}}(0,0)]^2 + 2K_2(0)k_2^{\text{ETH}}(0,0) \\ &\quad + \frac{1}{D} \sum_{i \neq j \neq l} K_4(\omega_{ij}, -\omega_{ij}, \omega_{il}) \\ &\quad + \frac{1}{D} \sum_{i \neq j} K_2(\omega_{ij})K_2(-\omega_{ij}). \end{aligned} \quad (\text{G12})$$

The fourth-order thermal free cumulant is

$$k_4(t, t, 0, 0) = C_4(t) - [k_2(0, 0)]^2 - [k_2(t, 0)]^2. \quad (\text{G13})$$

From Eq. (G6), at equal times one has

$$[k_2(0, 0)]^2 \simeq [k_2^{\text{ETH}}(0, 0)]^2 + 2K_2(0)k_2^{\text{ETH}}(0, 0), \quad (\text{G14})$$

moreover, under the non-resonance condition,

$$[[k_2(t, 0)]^2]_\infty = \mathcal{O}(D^{-2}), \quad (\text{G15})$$

so that this term does not contribute at order D^{-1} . Combining Eqs. (G12), (G13), and (G14), we obtain

$$\begin{aligned} [k_4(t, t, 0, 0)]_\infty &\simeq \frac{1}{D} \sum_{i \neq j \neq l} K_4(\omega_{ij}, -\omega_{ij}, \omega_{il}) \\ &\quad + \frac{1}{D} \sum_{i \neq j} K_2(\omega_{ij})K_2(-\omega_{ij}). \end{aligned} \quad (\text{G16})$$

The long-time plateau of the fourth-order thermal free cumulant is therefore determined, at the first non-vanishing order, by the subleading corrections and by the crossing term. Since $K_2 \sim D^{-1}$ and $K_4 \sim D^{-3}$, both terms scale as D^{-1} .

Therefore, the subleading corrections derived in this work are responsible for the finite-size late-time plateaus, which are exponentially suppressed in the system size. This provides a quantitative characterization of long-time freeness in chaotic dynamics.

-
- [1] A. Peres, *Ergodicity and mixing in quantum theory. I*, *Phys. Rev. A* **30**, 504 (1984).
 - [2] M. Feingold and A. Peres, *Distribution of matrix elements of chaotic systems*, *Phys. Rev. A* **34**, 591 (1986).
 - [3] M. Srednicki, *Chaos and quantum thermalization*, *Phys. Rev. E* **50**, 888 (1994).
 - [4] M. Srednicki, *The approach to thermal equilibrium in quantized chaotic systems*, *J. Phys. A* **32**, 1163 (1999).
 - [5] J. M. Deutsch, *Quantum statistical mechanics in a closed system*, *Phys. Rev. A* **43**, 2046 (1991).
 - [6] P. Reimann, *Eigenstate thermalization: Deutsch's approach and beyond*, *New J. Phys.* **17**, 055025 (2015).
 - [7] L. D'Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, *From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics*, *Adv. Phys.* **65**, 239 (2016).
 - [8] O. Bohigas, M. J. Giannoni, and C. Schmit, *Characterization of chaotic quantum spectra and universality of level fluctuation laws*, *Phys. Rev. Lett.* **52**, 1 (1984).
 - [9] M. V. Berry, *Regular and irregular semiclassical wavefunctions*, *J. Phys. A* **10**, 2083 (1977).
 - [10] A. Voros, *Semi-classical ergodicity of quantum eigenstates in the Wigner representation*, *Stochastic Behavior in Classical and Quantum Hamiltonian Systems*, Lecture Notes in Physics, edited by G. Casati-J. Ford (Springer, Berlin, Heidelberg, 1979), Vol. 93, pp. 326–333, [10.1007/BFb0021757](#).
 - [11] S. Goldstein, J. L. Lebowitz, R. Tumulka, and N. Zanghì, *Canonical typicality*, *Phys. Rev. Lett.* **96**, 050403 (2006).

- [12] S. Popescu, A. J. Short, and A. Winter, *Entanglement and the foundations of statistical mechanics*, *Nat. Phys.* **2**, 754 (2006).
- [13] M. Rigol, V. Dunjko, and M. Olshanii, *Thermalization and its mechanism for generic isolated quantum systems*, *Nature (London)* **452**, 854 (2008).
- [14] J. von Neumann, *Proof of the ergodic theorem and the H-theorem in quantum mechanics*, *Eur. Phys. J. H* **35**, 201 (2010).
- [15] S. Goldstein, J. L. Lebowitz, C. Mastrodonato, R. Tumulka, and N. Zanghi, *Approach to thermal equilibrium of macroscopic quantum systems*, *Phys. Rev. E* **81**, 011109 (2010).
- [16] A. Dymarsky, N. Lashkari, and H. Liu, *Subsystem eigenstate thermalization hypothesis*, *Phys. Rev. E* **97**, 012140 (2018).
- [17] L. Foini and J. Kurchan, *Eigenstate thermalization hypothesis and out of time order correlators*, *Phys. Rev. E* **99**, 042139 (2019).
- [18] A. Chan, A. De Luca, and J. T. Chalker, *Eigenstate correlations, thermalization, and the butterfly effect*, *Phys. Rev. Lett.* **122**, 220601 (2019).
- [19] C. Murthy and M. Srednicki, *Bounds on chaos from the eigenstate thermalization hypothesis*, *Phys. Rev. Lett.* **123**, 230606 (2019).
- [20] G. O. Alves, F. Fritzsche, and P. W. Claeys, *Probes of full eigenstate thermalization in ergodicity-breaking quantum circuits*, *Quantum* **9**, 1949 (2025).
- [21] J. Richter, A. Dymarsky, R. Steinigeweg, and J. Gemmer, *Eigenstate thermalization hypothesis beyond standard indicators: Emergence of random-matrix behavior at small frequencies*, *Phys. Rev. E* **102**, 042127 (2020).
- [22] D. Hahn, D. J. Luitz, and J. T. Chalker, *Eigenstate correlations, the eigenstate thermalization hypothesis, and quantum information dynamics in chaotic many-body quantum systems*, *Phys. Rev. X* **14**, 031029 (2024).
- [23] L. Foini, A. Dymarsky, and S. Pappalardi, *Out-of-equilibrium eigenstate thermalization hypothesis*, *Scipost Phys.* **18**, 136 (2025).
- [24] A. Belin and J. de Boer, *Random statistics of OPE coefficients and Euclidean wormholes*, *Classical Quantum Gravity* **38**, 164001 (2021).
- [25] A. Belin, J. de Boer, and D. Liska, *Non-Gaussianities in the statistical distribution of heavy OPE coefficients and wormholes*, *J. High Energy Phys.* **06**, 116 (2022).
- [26] J. Chandra, S. Collier, T. Hartman, and A. Maloney, *Semiclassical 3D gravity as an average of large- c CFTs*, *J. High Energy Phys.* (2022), 069 (2022).
- [27] D. L. Jafferis, D. K. Kolchmeyer, B. Mukhametzhanov, and J. Sonner, *JT gravity with matter, generalized ETH, and random matrices*, *Phys. Rev. D* **108**, 066015 (2023).
- [28] P. Hosur, X.-L. Qi, D. A. Roberts, and B. Yoshida, *Chaos in quantum channels*, *J. High Energy Phys.* **2016**, 004 (2016).
- [29] T. Kawamoto, *A strategy for proving the strong eigenstate thermalization hypothesis: Chaotic systems and holography*, *J. High Energy Phys.* **2025**, 095 (2025).
- [30] P. Saad, *Late time correlation functions, baby universes, and ETH in JT gravity*, *arXiv:1910.10311*.
- [31] D. L. Jafferis, D. K. Kolchmeyer, B. Mukhametzhanov, and J. Sonner, *Matrix models for eigenstate thermalization*, *Phys. Rev. X* **13**, 031033 (2023).
- [32] S. Pappalardi, L. Foini, and J. Kurchan, *Eigenstate thermalization hypothesis and free probability*, *Phys. Rev. Lett.* **129**, 170603 (2022).
- [33] S. Pappalardi, F. Fritzsche, and Tž Prosen, *Full eigenstate thermalization via free cumulants in quantum lattice systems*, *Phys. Rev. Lett.* **134**, 140404 (2025).
- [34] D.-V. Voiculescu, *Limit laws for Random matrices and free products*, *Inventiones Mathematicae* **104**, 201 (1991).
- [35] D.-V. Voiculescu, K. J. Dykema, and A. Nica, *Free Random Variables*, CRM Monograph Series, Vol. 1 (American Mathematical Society, Providence, RI, 1992).
- [36] R. Speicher, *Free probability theory and non-crossing partitions*, *Séminaire Lotharingien de Combinatoire* **39**, B39c (1997), <https://www.emis.de/journals/SLC/wpapers/s39speicher.pdf>.
- [37] H. A. Camargo, Y. Fu, V. Jahnke, K. Pal, and K.-Y. Kim, *Quantum signatures of chaos from free probability*, *J. High Energy Phys.* **2025**, 138 (2025).
- [38] F. Fritzsche and P. W. Claeys, *Free probability in a minimal quantum circuit model*, *arXiv:2506.11197*.
- [39] F. Fritzsche, Tž Prosen, and S. Pappalardi, *Microcanonical free cumulants in lattice systems*, *Phys. Rev. B* **111**, 054303 (2025).
- [40] F. Fritzsche, G. O. Alves, M. A. Rampp, and P. W. Claeys, *Free cumulants and full eigenstate thermalization from boundary scrambling*, *arXiv:2509.08060*.
- [41] T. Pathak, *Full eigenstate thermalization in integrable spin systems*, *arXiv:2510.05887*.
- [42] S. Jindal and P. Hosur, *Generalized free cumulants for quantum chaotic systems*, *J. High Energy Phys.* **2024**, 066 (2024).
- [43] M. Fava, J. Kurchan, and S. Pappalardi, *Designs via free probability*, *Phys. Rev. X* **15**, 011031 (2025).
- [44] G. Cipolloni, László Erdős, and D. Schröder, *Thermalisation for Wigner matrices*, *J. Funct. Anal.* **282**, 109394 (2022).
- [45] H. J. Chen and J. Kudler-Flam, *Free Independence and the noncrossing partition lattice in dual-unitary quantum circuits*, *Phys. Rev. B* **111**, 014311 (2025).
- [46] E. Vallini and S. Pappalardi, *Long-time freeness in the kicked top*, *Quantum* **10**, 2129 (2026).
- [47] N. Dowling, J. De Nardis, M. Heinrich, X. Turkeshi, and S. Pappalardi, *Free independence and unitary design from random matrix product unitaries*, *arXiv:2508.00051*.
- [48] L. Foini and J. Kurchan, *Eigenstate thermalization and rotational invariance in ergodic quantum systems*, *Phys. Rev. Lett.* **123**, 260601 (2019).
- [49] In this work we focus mostly on random unitary rotations, that is, when the system belongs to the unitary symmetry class. The orthogonal case is only briefly mentioned in Appendixes E and F 2.
- [50] E. P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, *Ann. Math.* **62**, 548 (1955).
- [51] M. L. Mehta, *Random Matrices*, 3rd ed. (Elsevier, Amsterdam, 2004).
- [52] M. Brenes, J. Goold, and M. Rigol, *Low-frequency behavior of off-diagonal matrix elements in the integrable*

- XXZ chain and in a locally perturbed quantum-chaotic XXZ chain, *Phys. Rev. B* **102**, 075127 (2020).
- [53] J. Wang, M. H. Lamann, J. Richter, R. Steinigeweg, A. Dymarsky, and J. Gemmer, *Eigenstate thermalization hypothesis and its deviations from random-matrix theory beyond the thermalization time*, *Phys. Rev. Lett.* **128**, 180601 (2022).
- [54] C. Bertoni, C. Wassner, G. Guarnieri, and J. Eisert, *Typical thermalization of low-entanglement states*, *Commun. Phys.* **8**, 301 (2025).
- [55] J. Wang, J. Richter, M. H. Lamann, R. Steinigeweg, J. Gemmer, and A. Dymarsky, *Emergence of unitary symmetry of microcanonically truncated operators in chaotic quantum systems*, *Phys. Rev. E* **110**, L032203 (2024).
- [56] J. Wang, H. Yan, R. Steinigeweg, and J. Gemmer, *Eigenstate thermalization hypothesis and random matrix theory universality in few-body systems*, [arXiv:2506.09011](https://arxiv.org/abs/2506.09011).
- [57] M. Füllgraf, J. Gemmer, and J. Wang, *Scaling of free cumulants in closed system-bath setups*, *SciPost Phys. Core* **9**, 025 (2026).
- [58] G. Casati, L. Molinari, and F. Izrailev, *Scaling properties of band random matrices*, *Phys. Rev. Lett.* **64**, 1851 (1990).
- [59] Y. Fyodorov and A. Mirlin, *Statistical properties of random banded matrices with sparse structure*, *Phys. Rev. Lett.* **67**, 2049 (1991).
- [60] D. Bernard and L. Hruza, *Structured random matrices and cyclic cumulants: A free probability approach*, *Random Matrices Theory Appl.* **13**, 2450014 (2024).
- [61] B. Collins and P. Śniady, *Integration with respect to the Haar measure on unitary, orthogonal and symplectic group*, *Commun. Math. Phys.* **264**, 773 (2006).
- [62] E. Brézin, C. Itzykson, G. Parisi, and J. B. Zuber, *Commun. Math. Phys.* **59**, 35 (1978).
- [63] P. Cvitanović, *Planar perturbation expansion*, *Phys. Lett. B* **99**, 49 (1981).
- [64] P. Cvitanović, P. G. Lauwers, and P. N. Scharbach, *The planar sector of field theories*, *Nucl. Phys. B* **203**, 385 (1982).
- [65] B. Collins and I. Nechita, *Random matrix techniques in quantum information theory*, *J. Math. Phys.* **57**, 015215 (2016).
- [66] B. Collins and I. Nechita, *Random quantum channels I: Graphical calculus and the Bell state phenomenon*, *Commun. Math. Phys.* **297**, 345 (2010).
- [67] J. Kudler-Flam, V. Narovlansky, and S. Ryu, *Distinguishing random and black hole microstates*, *PRX Quantum* **2**, 040340 (2021).
- [68] J. Kudler-Flam, V. Narovlansky, and S. Ryu, *Negativity spectra in random tensor networks and holography*, *J. High Energy Phys.* **2022**, 076 (2022).
- [69] N. Cheng, C. Lancien, G. Penington, M. Walter, and F. Witteveen, *Random tensor networks with non-trivial links*, *Ann. Henri Poincaré* **25**, 2107 (2024).
- [70] R. Movassagh and A. Edelman, *Isotropic entanglement*, [arXiv:1012.5039](https://arxiv.org/abs/1012.5039).
- [71] R. Movassagh and A. Edelman, *Density of states of quantum spin systems from isotropic entanglement*, *Phys. Rev. Lett.* **107**, 097205 (2011).
- [72] J. Chen, E. Hontz, J. Moix, M. Welborn, T. Van Voorhis, A. Suárez, R. Movassagh, and A. Edelman, *Error analysis of free probability approximations to the density of states of disordered systems*, *Phys. Rev. Lett.* **109**, 036403 (2012).
- [73] L. Hruza and D. Bernard, *Coherent fluctuations in noisy mesoscopic systems, the open quantum SSEP, and free probability*, *Phys. Rev. X* **13**, 011045 (2023).
- [74] M. Bauer, D. Bernard, P. Biane, and L. Hruza, *Bernoulli variables, classical exclusion processes and free probability*, *Ann. Henri Poincaré* **1** (2023).
- [75] D. Bernard and L. Hruza, *Out-of-equilibrium eigenstate thermalization hypothesis*, *SciPost Phys.* **15**, 175 (2023).
- [76] M. Berkooz, M. Isachenkov, V. Narovlansky, and G. Torrents, *Towards a full solution of the large N double-scaled SYK model*, *J. High Energy Phys.* **2019**, 079 (2019).
- [77] G. Penington, S. H. Shenker, D. Stanford, and Z. Yang, *Replica wormholes and the black hole interior*, *J. High Energy Phys.* **2022**, 205 (2022).
- [78] J. Wang, *Beyond islands: A free probabilistic approach*, *J. High Energy Phys.* **2023**, 040 (2023).
- [79] S. Wu, *Non-commutative probability insights into the double-scaling limit SYK model with constant perturbations: moments, cumulants and q-independence*, *J. Phys. A* **57**, 325203 (2024).
- [80] V. Chandrasekaran, G. Penington, and E. Witten, *Large N algebras and generalized entropy*, *J. High Energy Phys.* **2023**, 009 (2023).
- [81] R. Speicher, *Free probability theory: And its avatars in representation theory, random matrices, and operator algebras; also featuring: Non-commutative distributions, Jahresbericht der Deutschen Mathematiker-Vereinigung* **119**, 3 (2017).
- [82] J. A. Mingo and R. Speicher, *Free Probability and Random Matrices* (Springer, 2017), Vol. 35.
- [83] X.-G. Xia, *A simple introduction to free probability theory and its application to random matrices*, [arXiv:1902.10763](https://arxiv.org/abs/1902.10763).
- [84] These averages of products of matrix elements are the only ones formulated by the full ETH [see Eq. (1)]; the interest towards those quantities is motivated by the evaluation of thermal correlation functions. Indeed, when calculating thermal moments, they naturally appear:
- $$\langle A^n \rangle_\beta = \frac{\text{Tr}(A^n e^{-\beta H})}{\text{Tr}(e^{-\beta H})} = \frac{1}{Z} \sum_{i_1, \dots, i_n} A_{i_1 i_2} A_{i_2 i_3} \cdots A_{i_n i_1} e^{-\beta E_{i_1}}, \quad (42)$$
- with $Z = \sum_i e^{-\beta E_i}$ partition function at temperature $T = \beta^{-1}$.
- [85] Those products of matrix elements are usually excluded by ETH discussions, because they give a negligible contribution in the calculation of thermal correlation functions [see Eq. (42)], for entropic reasons. This also explains why free probability, rather than classical probability theory, is relevant for quantum dynamics.
- [86] A. Maillard, L. Foini, A. L. Castellanos, F. Krzakala, M. Mézard, and L. Zdeborová, *High-temperature expansions and message passing algorithms*, *J. Stat. Mech.* **2019**, 113301 (2019).
- [87] From the derivation in Appendix B, it becomes clear that higher-order contributions are considerably more difficult to compute, which goes beyond the scope of this work. Therefore, we do not provide subleading corrections for simple loops.

- [88] G. Barraquand and D. Bernard, *Introduction to quantum exclusion processes*, [arXiv:2507.01570](#).
- [89] R. Mondaini and M. Rigol, *Eigenstate thermalization in the two-dimensional transverse field Ising model. II. Off-diagonal matrix elements of observables*, *Phys. Rev. E* **96**, 012157 (2017).
- [90] M. Bukov, L. D'Alessio, and A. Polkovnikov, *Universal high-frequency behavior of periodically driven systems: From dynamical stabilization to Floquet engineering*, *Adv. Phys.* **64**, 139 (2015).
- [91] C. Fleckenstein and M. Bukov, *Thermalization and prethermalization in periodically kicked quantum spin chains*, *Phys. Rev. B* **103**, 144307 (2021).
- [92] M. Akila, D. Waltner, B. Gutkin, and T. Guhr, *Particle-time duality in the kicked Ising spin chain*, *J. Phys. A* **49**, 375101 (2016).
- [93] B. Bertini, P. Kos, and TŽ Prosen, *Exact spectral form factor in a minimal model of many-body quantum chaos*, *Phys. Rev. Lett.* **121**, 264101 (2018).
- [94] P. Braun, D. Waltner, M. Akila, B. Gutkin, and T. Guhr, *Transition from quantum chaos to localization in spin chains*, *Phys. Rev. E* **101**, 052201 (2020).
- [95] E. Vallini, L. Foini, and S. Pappalardi, *Codes and data for: "Refinements of the Eigenstate thermalization hypothesis under local rotational invariance via free probability"*, [10.5281/zenodo.17659462](#) (2025).
- [96] These interactions, similarly to those considered in [97], were specifically chosen to ensure the system belongs to the unitary symmetry class.
- [97] O. Bouverot-Dupuis, S. Pappalardi, J. Kurchan, A. Polkovnikov, and L. Foini, *Out-of-equilibrium eigenstate thermalization hypothesis*, *SciPost Phys.* **19**, 050 (2025).
- [98] F. Haake, *Quantum Signatures of Chaos*, 3rd ed. (Springer, Berlin, Heidelberg, 2010).
- [99] As these differences are also modulo 2π , we restrict to the interval $\omega_{\alpha\beta} \in [-\pi, \pi)$.
- [100] P. Weinberg and M. Bukov, *Out-of-equilibrium eigenstate thermalization hypothesis*, *SciPost Phys.* **2**, 003 (2017).
- [101] P. Weinberg, M. Bukov, and M. Schneider, *Quspin: A python package for quantum many-body dynamics, 2017-2025*, <https://github.com/weinbe58/QuSpin>.
- [102] Y. Zhang and P. Zhang, *Finite-size scaling of the full eigenstate thermalization in quantum spin chains*, [arXiv:2602.01809](#).
- [103] L. Vidmar and M. Rigol, *Generalized Gibbs ensemble in integrable lattice models*, *J. Stat. Mech.* **2016**, 064007 (2016).
- [104] M. Kliczkowski, RŁ Świątek, M. Hopjan, and L. Vidmar, *Fading ergodicity*, *Phys. Rev. B* **110**, 134206 (2024).
- [105] D. K. Mark, F. Surace, A. Elben, A. L. Shaw, J. Choi, G. Refael, M. Endres, and S. Choi, *Maximum entropy principle in deep thermalization and in Hilbert-space ergodicity*, *Phys. Rev. X* **14**, 041051 (2024).
- [106] M. McGinley and T. Schuster, *The scrooge ensemble in many-body quantum systems*, [arXiv:2511.17172](#).
- [107] S. Ghosh, C. M. Langlett, N. Hunter-Jones, and J. F. Rodriguez-Nieva, *Late-time ensembles of quantum states in quantum chaotic systems*, *Phys. Rev. B* **112**, 094302 (2025).
- [108] E. Tirrito, X. Turkeshi, and P. Sierant, *Anticoncentration and nonstabilizerness spreading under ergodic quantum dynamics*, *Phys. Rev. Lett.* **135**, 220401 (2025).
- [109] L. Cui, T. Schuster, L. Mao, H.-Y. Huang, and F. G. S. L. Brandão, *Random unitaries from Hamiltonian dynamics*, [arXiv:2510.08434](#).
- [110] This is not the case for the entire symmetric group; indeed, the entire Cayley graph itself does not form a lattice.
- [111] This is not true if considering the geodesic interval from the identity to a generic maximal cycle: in this case, it is an isomorphism, but not an embedding.
- [112] K. Kaneko, E. Iyoda, and T. Sagawa, *Characterizing complexity of many-body quantum dynamics by higher-order eigenstate thermalization*, *Phys. Rev. A* **101**, 042126 (2020).
- [113] A. Nica and R. Speicher, *Lectures on the Combinatorics of Free Probability*, London Mathematical Society Lecture Note Series, Vol. 335 (Cambridge University Press, Cambridge, UK, 2006).
- [114] This can be proven identifying S_n with an orientable surface, and applying Euler's formula: the only possible solutions to the triangle inequality are $d(e, \beta) + d(\beta, \alpha) = d(e, \alpha) + 2g$, for integer $g \geq 0$ [113].
- [115] The subtle point is that in general α is not forced to be in the interval $[e, \gamma_n]$, therefore one cannot identify it with a non-crossing partition; however, it is possible to construct another permutation $\alpha' = \lambda^{-1}\alpha\lambda$, with $\lambda \in S_n$, so that $\#\alpha = \#\alpha'$ and $\alpha' \in [e, \gamma_n]$. In this case $\mathcal{K}_\alpha = \mathcal{K}_{\alpha'}$: the term $\mathcal{K}$ is the same when taking α or α' and therefore can still be identified with the free cumulant associated to α' ; we will keep the notation α in the text, intending that when α is not in the interval $[e, \gamma_n]$, it has to be thought as α' .
- [116] The fact that a product of matrix elements can be described uniquely by a partition happens only in the case of ETH-cycles, which partitions the set $\mathbf{i}$ in block of equal indices, since one needs the indices to be on a circle for being in blocks of partitions.
- [117] Actually, the only reason why we can study the first order contribution is because of the power $1/D^2$ in Eq. (B6).
- [118] Harish-Chandra, *Differential operators on a semisimple Lie Algebra*, *Am. J. Math.* **79**, 87 (1957).
- [119] C. Itzykson and J.-B. Zuber, *The planar approximation. II*, *J. Math. Phys.* **21**, 411 (1980).
- [120] J. Bun, J.-P. Bouchaud, and M. Potters, *Cleaning large correlation matrices: Tools from random matrix theory*, *Phys. Rep.* **666**, 1 (2017).
- [121] E. Marinari, G. Parisi, and F. Ritort, *Replica field theory for deterministic models. II. A non-random spin glass with glassy behaviour*, *J. Phys. A* **27**, 7647 (1994).
- [122] A. Guionnet and M. Maïda, *A Fourier view on the R-transform and related asymptotics of spherical integrals*, *J. Funct. Anal.* **222**, 435 (2005).
- [123] For non-Floquet systems one should consider further corrections coming from saddle-point fluctuations.