

Arbitrary Polynomial Separations in Trainable Quantum Machine Learning

Eric R. Anschuetz^{1,2,3} and Xun Gao⁴

¹Institute for Quantum Information and Matter, Caltech, Pasadena, CA, USA

²Walter Burke Institute for Theoretical Physics, Caltech, Pasadena, CA, USA

³MIT Center for Theoretical Physics, Cambridge, MA, USA

⁴JILA and Department of Physics, CU Boulder, Boulder, CO, USA

Recent theoretical results in quantum machine learning have demonstrated a general trade-off between the expressive power of quantum neural networks (QNNs) and their trainability; as a corollary of these results, practical exponential separations in expressive power over classical machine learning models are believed to be infeasible as such QNNs take a time to train that is exponential in the model size. We here circumvent these negative results by constructing a hierarchy of efficiently trainable QNNs that exhibit unconditionally provable, polynomial memory separations of arbitrary constant degree over classical neural networks—including state-of-the-art models, such as Transformers—in performing a classical sequence modeling task. This construction is also computationally efficient, as each unit cell of the introduced class of QNNs only has constant gate complexity. We show that contextuality—informally, a quantitative notion of semantic ambiguity—is the source of the expressivity separation, suggesting that other learning tasks with this property may be a natural setting for the use of quantum learning algorithms.

1 Introduction

The field of quantum computing has made remarkable strides in recent years, with many quantum experiments demonstrating the potential for quantum devices to efficiently solve classically intractable problems [1–6]. These experiments have largely involved performing sampling tasks due to the fact that quantum devices naturally sample from distributions that are generally regarded as difficult for classical computers to sample from [7–9]. This observation has motivated recent study [10–15] into the feasibility of using quantum devices to perform a class of machine learning tasks known as *generative modeling*: given samples from some target distribution $p(\mathbf{y})$, can one efficiently learn and then sample from some model distribution $q(\mathbf{y})$ closely approximating $p(\mathbf{y})$ using a quantum algorithm? Furthermore, to determine whether the overhead of constructing a quantum device to implement such a quantum algorithm is necessary, one is also interested in the question of *quantum advantage*: are there any quantum-efficient learning tasks which *require* quantum resources to efficiently perform?

Recent theoretical analysis of quantum machine learning (QML) algorithms has determined that, indeed, such learning tasks do exist [16–21]. These results rely on reducing machine learning tasks to classically hard computational problems—such as the discrete logarithm—which are known to be easy to solve using quantum devices [22]. The associated learning algorithms are then problem-specific, i.e., require prior knowledge of the classically-hard problem used in the reduction. For this reason, while these constructions are theoretically useful for assessing the promise of quantum learning algorithms, it is unclear how to train such QML algorithms in real-world learning settings.

Eric R. Anschuetz: eans@caltech.edu

One might hope that these problem-specific learning algorithms could be replaced by a problem-agnostic algorithm such as gradient descent. This is possible in classical machine learning, where it is known that gradient descent efficiently optimizes typical loss functions even if in the worst case the loss can be nonconvex [23, 24]. Unfortunately, it is known that this is not the case in the quantum setting; generally, the training of QML models using problem-agnostic algorithms is difficult due to *quantum trainability barriers* (QTBs) such as the presence of “barren plateaus” (exponential loss function concentration) [25–29] or poor local minima in the loss landscape [30–32]. More recently, it has been shown that QTBs can only be avoided when the dynamics of the network are constrained to belong to a low-dimensional subspace of the full Hilbert space [32–34]. In many such scenarios efficient classical simulation algorithms exist for the quantum model, putting into question the practical utility of QML models [35–37]. Though there are still rigorous quantum-classical separations remaining in QML even with these restrictions [38–40], the separations are modest, making a practical experimental implementation showcasing an advantage unlikely due to the large overhead of quantum error correction [41].

We here for the first time balance considerations of (problem-agnostic) trainability, efficiency, and large quantum advantage by constructing a hierarchy of quantum neural networks (QNNs) called *k-hypergraph recurrent neural networks* (*k*-HRNNs) satisfying three simultaneous properties:

1. **Trainability.** *k*-HRNNs avoid QTBs.
2. **Efficiency.** Each unit cell of a *k*-HRNN has quantum gate complexity independent of the model size.
3. **Large quantum advantage.** *k*-HRNNs exhibit arbitrary polynomial memory advantages over *any* classical neural network in performing a certain generative modeling task over sequences.

Namely, we show that for every *k*-HRNN of model size n there exists a certain sequence learning task that can be performed by the model to zero error. We then show that no classical network with fewer than $\binom{n}{k} - 1$ neurons can perform this task to *any* finite cross entropy. Importantly, the classical networks for which we prove a memory lower bound include state-of-the-art models such as Transformers [42]. *k*-HRNNs achieve this balance of efficient trainability and high expressiveness by having a considerable amount of structure: the latent state of the network is constrained to be a *quantum hypergraph state* [43, 44]. We show that this structure is sufficient to prevent QTBs from occurring in the associated loss landscape of the network, while still giving the network enough freedom to maintain a large quantum-classical expressivity separation.

Our result is asymptotically optimal, as there exists a classical algorithm capable of simulating *k*-HRNNs using memory $\Theta(\binom{n}{k})$ [45]. We emphasize that while an arbitrary-polynomial time separation persists during inference, this classical simulation algorithm only exhibits a *cubic* time overhead during training due to the relative inefficiency of training on a quantum computer. Similar classical simulation methods were previously thought to inhibit any large quantum-classical separation in QML [35–37]. Our work manifests one important detail of these results: while this conventional wisdom is true during *training* (at least when the algebraic structure is known a priori [46]), large separations can be maintained during *inference*. As inference is by far the most expensive cost of deploying networks today [47, 48], our results are an important example of how large quantum-classical separations in machine learning can exist even though these efficient classical simulation algorithms are known.

Our separation is also interpretable: we show that the classical memory lower bound follows from the large amount of *contextuality*—in a technical sense [49, 50]—in the learning task we consider. Informally, the semantic meaning of any given token strongly depends on other tokens in the sequence. We prove that the presence of this contextuality inhibits efficient classical representations of the data. Though our result is proven on a specific data set, this suggests sequential data with a similar correlation structure may exhibit similar classical memory lower bounds. Indeed, contextuality has previously been used as a rigorous model for linguistic ambiguity [51], relational databases [52], and other settings where data has a “global inconsistency” structure [53, 54]. One interpretation of our results is then

that this correlation structure inhibits efficient classical representations of data but not necessarily efficient quantum representations of data, motivating problems exhibiting semantic ambiguity as natural settings in which to use QML algorithms.

Our results also have implications beyond machine learning; our separation can also be viewed as a quantum-classical communication separation in sampling from a certain conditional distribution $p(y_1, \dots, y_\ell \mid x_1, \dots, x_\ell)$, where ℓ parties each receive one x_i and output the associated y_i . Viewed through this lens, our results differ from previously known quantum-classical communication complexity separations [55–61] not only in the strength of the error model we consider, but also as the quantum algorithm capable of performing this task is efficient—each party only requires constant gate complexity. While there are many parties—the sequence lengths are on the order of $\binom{n}{k}$ in our learning task—this is natural in a streaming setting, where one hopes to minimize the per-token resource requirements in processing a long stream of tokens.

We proceed in the following way. In Sec. 2, we review basic concepts in quantum computation and neural sequence learning. Then, in Sec. 3, we introduce a new class of QNN models and demonstrate that it avoids quantum barriers to trainability. We then introduce a sequence learning task in Sec. 4, and give large classical communication-complexity lower-bounds in performing this task; simultaneously, we show that the QNN model we introduce is able to efficiently perform this task. Finally, we end in Sec. 5 with discussion on the implications of our results.

2 Preliminaries

2.1 Quantum Generative Models

We first review quantum generative models. Given a quantum state $|\psi\rangle$ on either an n -qubit or n -qumode system, there is a natural induced probability distribution given by the *measurement distribution*:

$$q(y_1, \dots, y_\ell) = \langle \psi | \Pi_{y_1} \dots \Pi_{y_\ell} | \psi \rangle, \quad (1)$$

where Π_{y_i} are Hermitian projectors (assumed here to be diagonal). More generally, one can consider a measurement distribution parameterized by $\theta \in \mathbb{R}^p$:

$$q_\theta(y_1, \dots, y_\ell) = \langle \psi | U_\theta^\dagger \Pi_{y_1} \dots \Pi_{y_\ell} U_\theta | \psi \rangle, \quad (2)$$

where U_θ is a parameterized unitary operator. When θ is optimized such that q_θ approximates some distribution of interest p , we say that $q_\theta(y_1, \dots, y_\ell)$ is a *quantum generative model*. On a quantum computer this optimization is typically performed by optimizing the maximum mean discrepancy (MMD) loss function, which can be written as minimizing the expectation value of some diagonal Hermitian operator $O_{\mathcal{D}}$ depending only on the training set $\mathcal{D}$ [62]:

$$\ell(\theta) = \langle \psi | U_\theta^\dagger O_{\mathcal{D}} U_\theta | \psi \rangle. \quad (3)$$

If the class U_θ is chosen generically, this optimization is provably hard due to the presence of *quantum trainability barriers* (QTB) [32–34]. Such barriers include the exponential concentration of typical loss functions over uniformly sampled θ (commonly known as the presence of “barren plateaus”), or a loss landscape dominated by poor local minima. QTBs are provably avoided when [32–34, 62]:

1. U_θ is constrained to belong to a poly(n)-dimensional Lie subgroup G of the unitary group, and
2. both $i|\psi\rangle\langle\psi|$ and $iO_{\mathcal{D}}$ belong to the algebra generating G .

Indeed, one can show that training such models via gradient descent takes time at least $\Omega(\dim(G))$ [33, 34].

Finally, we introduce some notation. We will often reference the X and Z Pauli operators:

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad (4)$$

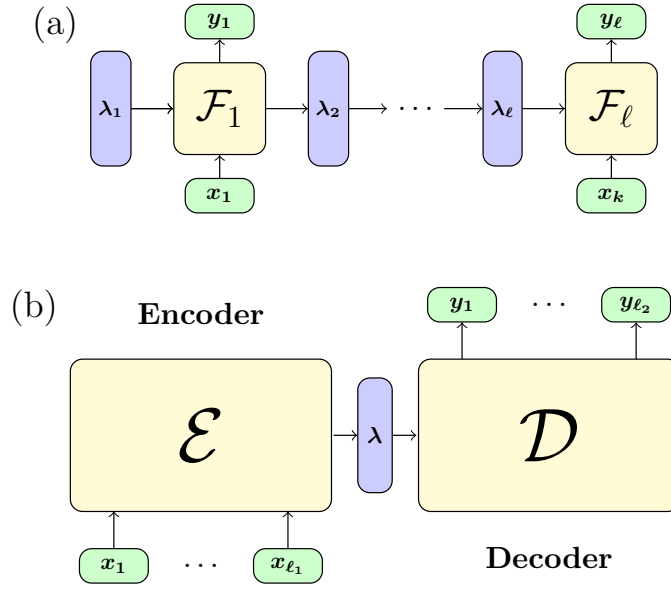


Figure 1: (a) An autoregressive neural sequence model. The model autoregressively takes input tokens x_i , and outputs decoded tokens y_i , with map $\mathcal{F}_i$. The model also has an unobserved internal memory with state $\lambda_i \in L$ after decoding the token x_{i-1} that $\mathcal{F}_i$ can depend on. (b) A general encoder-decoder model. $\mathcal{E}$ encodes the input $\mathbf{x}$ into some latent representation $\lambda \in L$. A decoder $\mathcal{D}$ then outputs the decoded sequence $\mathbf{y}$.

with subscript P_i shorthand for the n -fold tensor product $I_2 \otimes \dots \otimes P \otimes \dots \otimes I_2$ for Pauli operator P at the i th position. We will also consider the canonical position and momentum operators $\hat{q}_i, \hat{p}_i$ formally acting on an infinite-dimensional Hilbert space and obeying the so-called canonical commutation relation (in units where $\hbar = 1/2$):

$$[\hat{q}_j, \hat{p}_k] = \frac{i}{2} \delta_{jk}. \quad (5)$$

2.2 Neural Sequence Learning

Sequence-to-sequence or *sequence learning* [63] is the approximation of some given conditional distribution $p(\mathbf{y} | \mathbf{x})$ with a model distribution $q(\mathbf{y} | \mathbf{x})$, where the vectors $\mathbf{x}, \mathbf{y}$ are sequences of vectors $(x_1, \dots, x_{\ell_1}), (y_1, \dots, y_{\ell_2})$; we call each vector (x_i or y_i) a *token*. This framework encompasses sentence translation tasks [63], speech recognition [64], image captioning [65], and many more practical problems.

Sequence modeling today is typically performed using *neural sequence models*. Generally, these models are functions parameterized by some $\theta \in \mathbb{R}^p$ that take as input the sequence $\mathbf{x}$ and output a sample from the parameterized model distribution $q_\theta(\mathbf{y} | \mathbf{x})$. The parameters of these functions are trained to minimize an appropriate loss function, such as the (forward) empirical cross entropy:

$$\hat{H}(p, q | \theta) = -\frac{1}{|\mathcal{T}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{T}} p(\mathbf{y} | \mathbf{x}) \log(q_\theta(\mathbf{y} | \mathbf{x})), \quad (6)$$

where $\mathcal{T} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$ are samples from $p(\mathbf{x}, \mathbf{y})$. The backward empirical cross entropy is similarly defined, with $p \leftrightarrow q_\theta$.

One of the unique challenges when performing sequence modeling is the potentially large data dimensions due to data being comprised of arbitrarily long sequences. Thus, to maintain a resource scaling independent of the input sequence length, neural sequence models usually fall into one of

two classes: *autoregressive sequence models* [66–68] or *encoder-decoder models* [42, 63]. These classes of models are illustrated in Figs. 1(a) and 1(b), respectively; in the former, the trained network components are the unit cells $\mathcal{F}_i$, while in the latter the network is composed of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. Both models make use of a *latent vector* λ belonging to a *latent space* L in order to model long-range correlations between the tokens in a given data set. While the limit of large $\dim(\lambda)$ yields more expressive models, it also (typically) yields models less efficient to sample from. For this reason, we will use the dimension of λ as the metric by which we evaluate the efficiency of a given model.

We have not yet stated any requirements on the model functions $\mathcal{F}_i$, $\mathcal{E}$, or $\mathcal{D}$. If these functions were allowed to be fully general, they could in principle implement arbitrarily complicated, discontinuous functions. In practice, however, these $\mathcal{F}_i$, $\mathcal{E}$, and $\mathcal{D}$ are composed of elementary functions with some sort of smoothness structure such that efficient training of the network is possible. For this reason, we assume classical networks satisfy two properties on some finite-volume subspace $\mathcal{X}$ of the space of all inputs:

Assumption 2.1 (C^2 on $\mathcal{X}$). In the autoregressive setting, each $\mathcal{F}_i$ is continuously second-differentiable on $\mathcal{X}$. In the encoder-decoder setting, $\mathcal{E}$ is continuously second-differentiable on $\mathcal{X}$.

Assumption 2.2 (Strongly Morse on $\mathcal{X}$). In the autoregressive setting, the Jacobian of each $\mathcal{F}_i$ has determinant lower-bounded by a constant on $\mathcal{X}$. In the encoder-decoder setting, the Jacobian of $\mathcal{E}$ has determinant lower-bounded by a constant on $\mathcal{X}$.

Assumption 2.1 is straightforward and is automatically satisfied by models composed of linear layers along with rectified linear units, softmax functions, sigmoid functions, hyperbolic tangent functions, and other standard nonlinearities. Assumption 2.2 is slightly more subtle. Stated another way, we require that there is a finite-volume subspace of the space of inputs that are not near-critical points of the network. This is effectively a requirement that the model is *strongly Morse* [69] in a multivariate sense. Neural networks being strongly Morse is a common assumption [69–72], and certain classes of neural networks are known to be almost surely Morse over settings of their parameters [73]. In Appendix B we give other natural settings where these Assumptions are automatically satisfied.

3 Quantum k -Hypergraph Recurrent Neural Networks

We now introduce the hierarchy of quantum models that exhibit arbitrary polynomial expressivity separations over classical neural networks satisfying Assumptions 2.1 and 2.2. We call the k th level of this hierarchy of models *k -hypergraph recurrent neural networks* (k -HRNNs) for reasons that will soon become clear. We give a construction of these models in both a continuous variable (CV) setting and a qubit setting. In Appendix C, we also show that our k -HRNN construction has a natural interpretation as the quantization of a classical autoregressive architecture.

3.1 Qumode Construction

In a similar fashion to classical recurrent neural networks [66], k -HRNNs are autoregressive models—they are given input tokens $\mathbf{x}_i$ sequentially, and output tokens $\mathbf{y}_i$ sequentially. The model also has an n -qumode *latent state* $|\lambda\rangle$, generalizing the n -dimensional latent vector $\lambda \in \mathbb{R}^n$ in the classical case. We use $|\lambda_i\rangle$ to denote this state just prior to receiving the input token $\mathbf{x}_i$. We assume that the initial latent state is the position-squeezed state $|\lambda_1\rangle = |\hat{q} = 0\rangle^{\otimes n}$.

The trained components of the network are classical, given by four classical networks that are

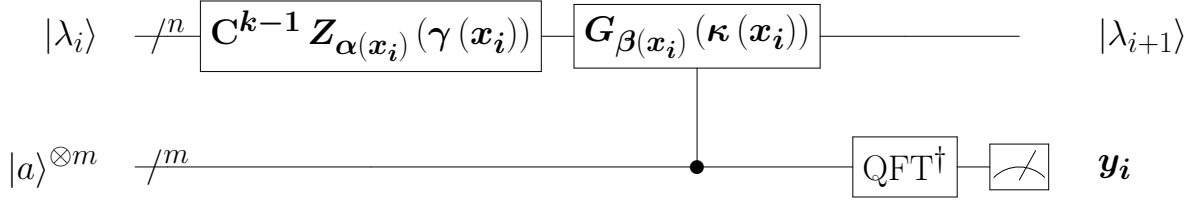


Figure 2: A k -hypergraph recurrent neural network (k -HRNN) unit cell as described in Sec. 3, with Greek letters denoting trained classical networks and G , as in Eq. (12). α and β output k -tuples of qumodes on which the associated operators have support. γ and κ output the associated rotation angles. The ancillary $|a\rangle^{\otimes m}$ and QFT perform phase estimation of the operators G , as described in Sec. 3. Measurement is either in the position basis (in the qumode setting) or in the computational basis (in the qubit setting).

functions of the input $\mathbf{x}_i$ at a given time step:

$$\alpha(\mathbf{x}_i) = (v_1, \dots, v_k) \in \binom{[n]}{k}; \quad (7)$$

$$\beta(\mathbf{x}_i) = (v_1, \dots, v_k) \in \binom{[n]}{k}; \quad (8)$$

$$\gamma(\mathbf{x}_i) \in \mathbb{R}; \quad (9)$$

$$\kappa(\mathbf{x}_i) = (\phi, \theta) \in \mathbb{R}^k \times \mathbb{R}^{2^k}. \quad (10)$$

Here, $\binom{[n]}{k}$ is the set of all sets of k distinct elements chosen from $\{1, \dots, n\}$. In what follows, we use the quantum gate:

$$C^{k-1} Z_{\bar{v}}(\gamma) = \exp\left(-i\gamma \prod_{i=1}^k \hat{q}_{v_i}\right), \quad (11)$$

i.e., the continuous-variable analogue of a multi-control Z -rotation acting on qumodes indexed by $\bar{v} \in \binom{[n]}{k}$. We also define:

$$G_{\bar{v}}(\phi, \theta) := \exp\left(-i \sum_{i=1}^k \phi_i \hat{p}_{v_i}\right) \exp\left(-i \sum_{\bar{w} \in 2^{\bar{v}}} \theta_{\bar{w}} \prod_{w_i \in \bar{w}} \hat{q}_{w_i}\right), \quad (12)$$

where here $2^{\bar{v}}$ denotes the power set of $\bar{v}$. For appropriate choices of ϕ and θ , $G_{\bar{v}}(\phi, \theta)$ can be interpreted as a stabilizer of a (*weighted*) *hypergraph state* [43], a class of states in one-to-one correspondence with weighted hypergraphs. Given a hypergraph with hyperedges $\bar{v} \in E$ and corresponding weights $w_{\bar{v}}$, the associated hypergraph state is:

$$|\psi\rangle = \prod_{\bar{v} \in E} C^{|\bar{v}|-1} Z_{\bar{v}}(w_{\bar{v}}) |\hat{p}=0\rangle^{\otimes n}. \quad (13)$$

In our setting, the maximal cardinality of $\bar{v} \in E$ is k . These states are reviewed in further detail in Appendix A.

We now describe the construction of a unit cell of a k -HRNN, where we assume for now that the dimension of each output token y_i is 1. First, the previously-described qumode analogue of a multi-control Z rotation is applied to the latent state of the model:

$$|\lambda'_i\rangle = C^{k-1} Z_{\alpha(\mathbf{x}_i)}(\gamma(\mathbf{x}_i)) |\lambda_i\rangle. \quad (14)$$

Next, phase estimation (modulo 2π) of the operator $G_{\beta(\mathbf{x}_i)}(\kappa(\mathbf{x}_i))$ is performed [74]. The measurement result y_i is the output of the cell, and the post-measurement state $|\lambda_{i+1}\rangle$ is the new latent state

of the model. The quantum circuit implementing this sequence of transformations is illustrated in Fig. 2, where $|a\rangle = |\text{GKP}\rangle$ denotes the position-squeezed *GKP state* [75]:

$$|\text{GKP}\rangle \propto \sum_{q=-\infty}^{\infty} |\hat{q} = \pi q\rangle \quad (15)$$

and QFT denotes the CV quantum Fourier transform $\mathcal{F}^{\otimes m}$. Higher-dimensional outputs $\mathbf{y}_i \in \mathbb{R}^m$ can be treated by performing m measurements of operators of the form of $G_{\bar{v}}(\phi, \theta)$ instead of just one. More concretely, one considers ordered sets of classical networks $(\beta_h(\mathbf{x}_i))_{h=1}^m$ and $(\kappa_h(\mathbf{x}_i))_{h=1}^m$, and sequentially performs phase estimation of the $G_{\beta_h(\mathbf{x}_i)}(\kappa_h(\mathbf{x}_i))$ to read out the elements of $\mathbf{y}_i$.

By construction, qumode k -HRNNs have evolution generated via the commutator by the operators $\{\hat{p}_i, \prod_{v_j \in \bar{v}} \hat{q}_{v_j}\}$, where $\bar{v}$ is of degree at most k . This set is closed under the commutator as:

$$\left[\hat{p}_i, \prod_{v_j \in \bar{v}} \hat{q}_{v_j} \right] = \frac{i}{2} \delta_{i \in \bar{v}} \prod_{v_j \neq i \in \bar{v}} \hat{q}_{v_j} \propto \prod_{w_j \in \bar{w} := \bar{v} \setminus \{i\}} \hat{q}_{w_j}. \quad (16)$$

Thus, for constant k , the dimension of the Lie subgroup G to which the dynamics are constrained is $\mathcal{O}(n^k)$. Furthermore, the initial state of the network is a stabilizer state that is stabilized by group elements generated by the $\hat{p}_i$. Finally, measurement is in the position basis, and the MMD loss discussed in Sec. 2.1 is a function of expectation values of low-degree polynomials of the $\hat{q}_i$ [62]. These facts taken together imply the absence of quantum trainability barriers (QTBs), as discussed in Sec. 2.1.

Indeed, QTBs can be avoided more directly with only a small overhead in time complexity during training. While training a k -HRNN on a quantum devices would take time at least $\Omega(n^k)$ (see Sec. 2.1), there exists a known classical algorithm for simulating Lie subgroup-constrained dynamics which, when applied to estimating the loss of a k -HRNN, achieves a time complexity of $\mathcal{O}(n^{2.8k})$ per unit cell [45]. For this reason, QTBs can be circumvented by avoiding training on a quantum device entirely, incurring only a cubic overhead in time complexity! We thus emphasize that the separation we discuss in Sec. 4 should be interpreted as an arbitrary-degree polynomial quantum-classical resource separation only during inference, i.e., after the model has been trained and deployed. We further discuss this subtlety in Appendix D.

3.2 Qubit Construction

The qubit construction is conceptually identical to the qumode construction, so we only highlight its differences. In the qubit setting, $|\lambda\rangle^{\otimes n}$ is an n -qubit quantum state initialized in the all-zero state $|\lambda_1\rangle = |0\rangle^{\otimes n}$, and the gates applied by the network are of the form:

$$C^{k-1} Z_{\bar{v}}(\gamma) = \exp(-i\gamma |0\rangle \langle 0|_{\bar{v}}), \quad (17)$$

$$G_{\bar{v}}(\phi, \theta) := \exp\left(-i \sum_{i=1}^k \phi_i X_{v_i}\right) \exp\left(-i \sum_{\bar{w} \in 2^{\bar{v}}} \theta_{\bar{w}} |1\rangle \langle 1|_{\bar{w}}\right), \quad (18)$$

where we have used the shorthand:

$$|1\rangle \langle 1|_{\bar{v}} := \prod_{v_i \in \bar{v}} \left(\frac{I_{2^n} - Z_{v_i}}{2} \right). \quad (19)$$

Phase estimation of $G_{\bar{v}}(\phi, \theta)$ can be performed using the same circuit as in the CV setting, diagrammed in Fig. 2. In practice, phase estimation can also be performed to finite precision using the standard qubit-based algorithm; that is, where $|a\rangle = |+\rangle$ and QFT denotes the qubit-based quantum Fourier transform [76]. All other aspects of the network are identical to the CV setting.

The main distinction between the qubit setting and the qumode setting is that, unlike the qumode setting, the subspace explored by the qubit k -HRNN is *not* low-dimensional. Indeed, in the qubit setting sampling from a k -HRNN is at least as difficult as sampling from the well-known instantaneous quantum polynomial-time (IQP) circuit class, widely believed to be exponentially difficult to classically simulate to small total variation distance even when $k = 2$ [77]. For this reason we cannot use the same arguments as in the qumode setting for the avoidance of QTBs in the training of the qubit-based network. In Appendix E we give a slightly more complicated construction of the qubit-based model which we argue is efficient to train, but no longer has the advantage of being independent of the token index i . Unfortunately, the model no longer being translationally invariant complicates practical implementations in a typical streaming setting when using this training method. One potential heuristic to circumvent this shortcoming might be to “round” the CV model—which can be efficiently trained classically in a translationally invariant setting, as discussed in Sec. 3.1—to a qubit model at constant precision. We hope to explore this approach further in future work.

4 Arbitrary Polynomial Expressivity Separations Over Classical Neural Networks

4.1 (ℓ, n, k) -Hypergraph Stabilizer Measurement Translation

Having introduced k -HRNNs, we now describe the classical sequence modeling task with which we will prove our separation. For clarity we here omit some technical details which are left for Appendix B.1.

We call the task we construct (ℓ, n, k) -hypergraph stabilizer measurement translation $((\ell, n, k)$ -HSMT), with both a qumode- and a qubit-based variant. Unless otherwise stated, we assume that k is constant with respect to n . For both variants of (ℓ, n, k) -HSMT, we consider ℓ -long input sequences

$$\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_\ell)^\top, \quad (20)$$

where

$$\mathbf{x}_i = (\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \gamma_i, \boldsymbol{\kappa}_i) \in [n]^k \times [n]^k \times \mathbb{R} \times (\mathbb{R}^k \times \mathbb{R}^{2^k}). \quad (21)$$

The task is to sample from the measurement distribution $p(y_1, \dots, y_\ell \mid \mathbf{x}_1, \dots, \mathbf{x}_\ell)$ of either a qumode- or a qubit-based k -HRNN, where the i th unit cell applies operations parameterized by—as described in Sec. 3— $(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \gamma_i, \boldsymbol{\kappa}_i)$. By construction, a k -HRNN can sample from this distribution exactly with n qumodes or qubits and a constant number of gates per unit cell.

4.2 Expressivity Separations

We now give memory lower bounds on classical neural networks performing (ℓ, n, k) -HSMT to a finite backward cross entropy whenever $k \geq 2$. We here only give a sketch of our arguments, leaving detailed proofs for Appendix B. Our main results follow from two facts that we prove:

Proposition 4.1. *Let $p(\mathbf{y} \mid \mathbf{x})$ be the distribution corresponding to the (ℓ, n, k) -HSMT task for some $\ell \geq \binom{n}{k} + n$. There exists a $\binom{n}{k} - 1$ -dimensional subspace $\mathcal{X}$ of inputs where the following holds: consider any triplet of distinct $\binom{n}{k}$ -token sequences $(\mathbf{x}, \mathbf{x}', \mathbf{x}'') \in \mathcal{X}^{\times 3}$. There exists some n -token sequence $\tilde{\mathbf{x}}$ such that, for all $\mathbf{y}$,*

$$p(\mathbf{y} \mid \mathbf{x} \oplus \tilde{\mathbf{x}}) \times p(\mathbf{y} \mid \mathbf{x}' \oplus \tilde{\mathbf{x}}) \times p(\mathbf{y} \mid \mathbf{x}'' \oplus \tilde{\mathbf{x}}) = 0. \quad (22)$$

Furthermore, this property is due to contextuality.

Proposition 4.2. *Any classical neural network satisfying Assumptions 2.1 and 2.2 with latent space dimension less than $\binom{n}{k} - 1$ maps some finite-volume subspace of $\mathcal{X}$ to a single point in latent space.*

Before proceeding with a proof sketch, we give some intuition behind these two Propositions—as well as their implications—by considering $k = 2$ as a concrete example.

Table 1: An example of quantum contextuality using a Mermin–Peres magic square [78]. All operators in each row and column commute. Additionally, the product of each row and column is the identity operator, except for the final column, which gives minus the identity. Thus, definite classical values cannot be assigned to each operator without yielding a contradiction.

X_1	X_2	$X_1 X_2$
$X_1 Z_2$	$Z_1 X_2$	$-X_1 Z_1 X_2 Z_2$
Z_2	Z_1	$Z_1 Z_2$

4.2.1 Intuition Behind Proposition 4.1

We begin with a simple example of Proposition 4.1 for $n = 2$ qubits. Recall that the $(\ell, 2, 2)$ -HSMT corresponds to simulating the dynamics of a 2-HRNN for ℓ time steps. We consider a triplet of 2-token input sequences $(\mathbf{x}, \mathbf{x}', \mathbf{x}'')$ which, when conditioned on the first two output tokens $y_1 = +1$ and $y_2 = +1$, respectively correspond to the preparation of three latent states:

$$|\lambda_2\rangle := |++\rangle, \quad (23)$$

$$|\lambda'_2\rangle := C Z_{1,2} |++\rangle, \quad (24)$$

$$|\lambda''_2\rangle := |00\rangle \quad (25)$$

in the associated k -HRNN. Stabilizers of these three states are written as the rows of Table 1. Note that the rows and columns all multiply to the identity matrix I outside of the third column, which multiplies to $-I$. This is a famous example of a *Mermin–Peres magic square* [78]: no classical assignment of these observables are consistent with the parity constraints enforced by their products, even though each row and column of operators are mutually commuting. This property is known as *contextuality*, which counterintuitively is known to be satisfiable in quantum theory [49, 50].

We now use this magic square to complete the sequences $(\mathbf{x}, \mathbf{x}', \mathbf{x}'')$ to three sequences with associated conditional distributions having infinite cross entropy. We consider a 2-token sequence $\tilde{\mathbf{x}}$ corresponding to the measurement of $Z_1 Z_2$ followed by $X_1 X_2$. From the definition of (ℓ, n, k) -HSMT, we have:

$$p(y_3 = +1, y_4 = -1 \mid \mathbf{x} \oplus \tilde{\mathbf{x}}, y_1 = +1, y_2 = +1) = 0; \quad (26)$$

$$p(y_3 = +1, y_4 = +1 \mid \mathbf{x}' \oplus \tilde{\mathbf{x}}, y_1 = +1, y_2 = +1) = 0; \quad (27)$$

$$p(y_3 = -1 \mid \mathbf{x}'' \oplus \tilde{\mathbf{x}}, y_1 = +1, y_2 = +1) = 0. \quad (28)$$

This is enough to show Eq. (22) when additionally conditioned on $y_1 = y_2 = +1$. Similar arguments for all other choices of y_1 and y_2 show this is indeed the case for all $\mathbf{y}$. This argument also generalizes for any choice of three states $|\lambda_2\rangle, |\lambda'_2\rangle, |\lambda''_2\rangle$, as long as they are stabilized by operators comprising the rows of a Mermin–Peres magic square.

For general k —as well as for qumodes—the proof of the Proposition proceeds similarly to the logic laid out here, though now the magic square is composed of operators in the $k - 1$ st level of the Clifford hierarchy [79] rather than Pauli operators. Formal proofs for qubits and qumodes are given in Appendix B.2 as the proofs of Lemmas B.1 and B.2, respectively.

4.2.2 Intuition Behind Proposition 4.2

We now sketch our proof of Proposition 4.2, which relies on Assumptions 2.1 and 2.2. At a basic level, this Proposition is a statement that if the classical neural network is “sufficiently nice,” $\mathcal{X}$ cannot be injectively represented in the latent space L of the network if $\dim(L) < \dim(\mathcal{X}) = \binom{n}{k} - 1$.

To demonstrate this we show that any such classical network exhibits a *fiber bundle* structure on some finite-volume subspace of inputs, where each fiber is “sufficiently large” such that it intersects $\mathcal{X}$ at three points (or more). We give a sketch of this structure in Fig. 3. Intuitively, this result can be

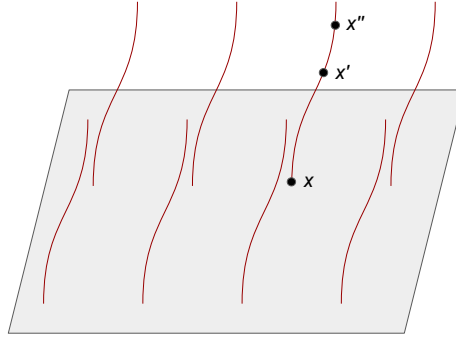


Figure 3: A representation of the noninjectivity of a classical network performing (ℓ, n, k) -HSMT as described in Proposition 4.2. On some finite-volume subspace of inputs, any classical network noninjectively maps all points on a red line (a “fiber”) to the same point in latent space, isomorphic to the gray parallelogram (the “base space”). $x, x', x'' \in \mathcal{X}$ label input sequences as described in Proposition 4.1. Due to the noninjectivity of the network, x , x' , and x'' are indistinguishable to the classical neural network.

seen as a consequence of the inverse function theorem, where Assumption 2.2 ensures that the theorem holds on some finite-volume subspace of inputs. More formally we use a so-called “global implicit function theorem” described in Ref. [80].

Formal proofs of this Proposition differ slightly depending on whether the classical model is an autoregressive model or an encoder-decoder model. These two cases are given as Theorems B.3 and B.4, respectively, both stated and proved in Appendix B.

4.2.3 Main Result

Our main result follows from Propositions 4.1 and 4.2. As previously stated, our formal proofs and statements depend on whether the classical model is an autoregressive model (stated as Theorem B.3) or an encoder-decoder model (stated as Theorem B.4), both stated and proved in Appendix B. For simplicity, we here gloss over the details which distinguish these two Theorems and give a high-level proof sketch of our separations.

Theorem 4.3 (Memory lower bound for (ℓ, n, k) -HSMT, informal). *Consider the qubit or qumode (ℓ, n, k) -HSMT task as described in Sec. 4.1 with $\ell \geq \binom{n}{k} + n$. A classical neural network satisfying Assumptions 2.1 and 2.2 must have latent space dimension at least $\binom{n}{k} - 1$ to perform the (ℓ, n, k) -HSMT task to any finite backward cross entropy. A k -HRNN on n qubits or qumodes and constant gate complexity per unit cell can perform this task to zero error.*

Proof sketch. We prove this statement by contradiction. Assume there exists a classical model of latent space dimension less than $\binom{n}{k} - 1$ that performs the (ℓ, n, k) -HSMT task to some finite backward cross entropy. Let $\mathcal{X}$ be as in Proposition 4.1. By Proposition 4.2, there exists some $\binom{n}{k}$ -token sequences $x, x', x'' \in \mathcal{X}$ that map to the same point in the latent space of the classical model. By Proposition 4.1, these inputs can be completed to ℓ -token sequences $x \oplus \tilde{x}$, $x' \oplus \tilde{x}$, and $x'' \oplus \tilde{x}$, where the associated conditional distributions in the (ℓ, n, k) -HSMT task have disjoint supports. However, as the model is unable to distinguish these inputs—as it maps these inputs to the same point in latent space—it is required to sample from some point in the intersection of their supports. This yields a contradiction. $\square$

Before concluding we note that this lower bound also holds when k scales with n , i.e., for $k = n/2$ this demonstrates an exponential classical memory lower bound for (ℓ, n, k) -HSMT. The trade-off is that the $k = n/2$ -HRNN which performs this task to zero error is not efficiently trainable by the discussion of Sec. 2.1, though this separation may be of independent interest as an exponential

quantum-classical communication complexity separation where the parties in the quantum setting have identical strategies implemented using only a constant number of gates.

5 Discussion

Our results construct a hierarchy quantum neural networks that not only avoids the trainability barriers typically encountered in quantum machine learning models, but also more expressive than classical neural networks by an arbitrary polynomial factor. We achieve this by considering a certain sequence learning task exhibiting a large amount of contextuality, and then demonstrating that the presence of this contextuality leads to a classical memory lower-bound in performing the task. Though the sequence lengths needed to witness a separation are large—on the order of n^k —this is natural in a streaming setting, where one wishes to perform sequence modeling in a way that minimizes the resources required to perform inference per token from an arbitrarily-long sequence.

Our introduced hierarchy of quantum models tracts nicely with experimental capabilities: given the ability to implement k -local gates, one can achieve a $\Theta\left(\binom{n}{k}\right)$ -memory separation over classical neural networks in performing the translation task we here introduce. This makes our hierarchy a natural fit for implementation on Rydberg atom-based quantum architectures which naturally implement multi-control Z gates [81]. Our model also has the benefit of being composed of operations for which error correction has successfully been experimentally demonstrated [82]. Cavity quantum electrodynamical (QED) systems are also a natural platform to implement our model, as coherent information from the CV degrees of freedom can be stored in the cavity while atomic degrees of freedom are read out to perform the necessary measurements [83, 84].

More work is required for demonstrating the practicality of our introduced advantage. First, qubit-based k -HRNNs are only efficiently trainable in a setting where the model is not translationally invariant; though this matters less in an offline learning scenario, it is unclear how a qubit k -HRNN would perform inference in a purely online setting. An alternative approach could be to consider an encoder-decoder variant of the qubit k -HRNN where translational invariance is maintained in both the encoder and decoder, though not together. Second, our protocol is here considered in a noise-free context, far from the experimental capabilities of most current quantum devices. However, our results follow from measurement sequences similar to those seen in the setting of quantum pseudo-telepathy games [85], and previous work [86, 87] has demonstrated instances where quantum separations based on pseudo-telepathy games can be made noise resilient. We hope to explore these connections in future work.

Furthermore, it is not yet clear how commonly contextuality—the specific correlational structure with which we prove our separation—appears in classical data. Concurrent work [88] has begun to investigate this, showing that standard machine learning data sets do possess nontrivial contextual correlations and that this inhibits classical representations of the data. Of course, this does not imply that learning such data automatically yields a “quantum advantage” regarding model size. Quantum easiness is the more difficult side of the argument; outside of constructed settings such as ours, it is difficult to analytically show that a data set can be efficiently represented quantumly. Our result can be interpreted as a statement that semantic ambiguity measured as nontrivial contextuality inhibits classical representations of data but does not necessarily inhibit quantum ones, pointing toward data with this correlational structure as promising empirical testing grounds for quantum machine learning.

HRNNs demonstrate that quantizing a very simple class of recurrent neural networks is enough to achieve a large expressivity separation over classical neural networks on a sequence modeling task. Given the tremendous progress in the performance of large language models in recent years with the introduction of, for instance, GPT-3 [89] and GPT-4 [90], this begs the natural question: what can be achieved by quantizing more sophisticated classical models? We hope in the future to address this.

Acknowledgements

We thank Cameron Calcluth and Mikhail D. Lukin for helpful discussion. E.R.A. was funded in part by the Walter Burke Institute for Theoretical Physics at Caltech and in part by the DARPA ONISQ program (grant number W911NF2010021). X.G. acknowledges support from the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Systems Accelerator, NSF PFC grant No. PHYS 2317149, and start-up grants from CU Boulder.

References

- [1] F. Arute *et al.*, “Quantum supremacy using a programmable superconducting processor,” *Nature* **574**, 505 (2019).
- [2] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, *et al.*, “Quantum computational advantage using photons,” *Science* **370**, 1460 (2020).
- [3] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, *et al.*, “Strong quantum computational advantage using a superconducting quantum processor,” *Phys. Rev. Lett.* **127**, 180501 (2021).
- [4] H.-S. Zhong, Y.-H. Deng, J. Qin, H. Wang, M.-C. Chen, L.-C. Peng, Y.-H. Luo, D. Wu, S.-Q. Gong, H. Su, *et al.*, “Phase-programmable Gaussian boson sampling using stimulated squeezed light,” *Phys. Rev. Lett.* **127**, 180502 (2021).
- [5] Q. Zhu, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, M. Gong, *et al.*, “Quantum computational advantage via 60-qubit 24-cycle random circuit sampling,” *Sci. Bull.* **67**, 240 (2022).
- [6] D. Hangleiter and J. Eisert, “Computational advantage of quantum random sampling,” (2022), [arXiv:2206.04079 \[quant-ph\]](https://arxiv.org/abs/2206.04079).
- [7] S. Aaronson and A. Arkhipov, “The computational complexity of linear optics,” in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC ’11 (Association for Computing Machinery, New York, NY, USA, 2011) pp. 333–342.
- [8] A. Bouland, B. Fefferman, C. Nirkhe, and U. Vazirani, “On the complexity and verification of quantum random circuit sampling,” *Nat. Phys.* **15**, 159 (2019).
- [9] S. Aaronson and S. Gunn, “On the classical hardness of spoofing linear cross-entropy benchmarking,” *Theory Comput.* **16**, 1 (2020).
- [10] M. Schuld, I. Sinayskiy, and F. Petruccione, “An introduction to quantum machine learning,” *Contemp. Phys.* **56**, 172 (2015).
- [11] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature* **549**, 195 (2017).
- [12] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, “Quantum Boltzmann machine,” *Phys. Rev. X* **8**, 021050 (2018).
- [13] S. Lloyd and C. Weedbrook, “Quantum generative adversarial learning,” *Phys. Rev. Lett.* **121**, 040502 (2018).
- [14] M. Schuld and N. Killoran, “Quantum machine learning in feature Hilbert spaces,” *Phys. Rev. Lett.* **122**, 040504 (2019).
- [15] N. Killoran, T. R. Bromley, J. M. Arrazola, M. Schuld, N. Quesada, and S. Lloyd, “Continuous-variable quantum neural networks,” *Phys. Rev. Research* **1**, 033063 (2019).
- [16] X. Gao, Z.-Y. Zhang, and L.-M. Duan, “A quantum machine learning algorithm based on generative models,” *Sci. Adv.* **4**, eaat9004 (2018).
- [17] B. Coyle, D. Mills, V. Danos, and E. Kashefi, “The Born supremacy: quantum advantage and training of an Ising Born machine,” *npj Quantum Inf.* **6**, 1 (2020).
- [18] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, “Expressive power of parametrized quantum circuits,” *Phys. Rev. Research* **2**, 033125 (2020).

- [19] R. Sweke, J.-P. Seifert, D. Hangleiter, and J. Eisert, “On the Quantum versus Classical Learnability of Discrete Distributions,” *Quantum* **5**, 417 (2021).
- [20] Y. Liu, S. Arunachalam, and K. Temme, “A rigorous and robust quantum speed-up in supervised machine learning,” *Nat. Phys.* **17**, 1013 (2021).
- [21] E. Gil-Fuster, C. Gyurik, A. Pérez-Salinas, and V. Dunjko, “On the relation between trainability and dequantization of variational quantum learning models,” (2024), [arXiv:2406.07072 \[quant-ph\]](#).
- [22] P. W. Shor, “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer,” *SIAM J. Comput.* **26**, 1484 (1997).
- [23] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, “The loss surfaces of multilayer networks,” in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 38, edited by G. Lebanon and S. V. N. Vishwanathan (PMLR, San Diego, CA, USA, 2015) pp. 192–204.
- [24] P. A. Chaudhari, *A Picture of the Energy Landscape of Deep Neural Networks*, Ph.D. thesis, University of California, Los Angeles (2018).
- [25] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, “Barren plateaus in quantum neural network training landscapes,” *Nat. Commun.* **9**, 4812 (2018).
- [26] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, “Cost function dependent barren plateaus in shallow parametrized quantum circuits,” *Nat. Commun.* **12**, 1791 (2021).
- [27] J. Napp, “Quantifying the barren plateau phenomenon for a model of unstructured variational ansätze,” (2022), [arXiv:2203.06174 \[quant-ph\]](#).
- [28] M. Cerezo and P. J. Coles, “Higher order derivatives of quantum neural networks with barren plateaus,” *Quantum Sci. Technol.* **6**, 035006 (2021).
- [29] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, “Effect of barren plateaus on gradient-free optimization,” *Quantum* **5**, 558 (2021).
- [30] E. R. Anschuetz, “Critical points in quantum generative models,” in *International Conference on Learning Representations*, edited by K. Hofmann, A. Rush, Y. Liu, C. Finn, Y. Choi, and M. Deisenroth (OpenReview, 2022).
- [31] E. R. Anschuetz and B. T. Kiani, “Quantum variational algorithms are swamped with traps,” *Nat. Commun.* **13**, 7760 (2022).
- [32] E. R. Anschuetz, “A unified theory of quantum neural network loss landscapes,” (2024), [arXiv:2408.11901 \[quant-ph\]](#).
- [33] E. Fontana, D. Herman, S. Chakrabarti, N. Kumar, R. Yalovetzky, J. Heredge, S. H. Sureshbabu, and M. Pistoia, “The adjoint is all you need: Characterizing barren plateaus in quantum ansätze,” (2023), [arXiv:2309.07902 \[quant-ph\]](#).
- [34] M. Ragone, B. N. Bakalov, F. Sauvage, A. F. Kemper, C. O. Marrero, M. Larocca, and M. Cerezo, “A unified theory of barren plateaus for deep parametrized quantum circuits,” (2023), [arXiv:2309.09342 \[quant-ph\]](#).
- [35] E. R. Anschuetz, A. Bauer, B. T. Kiani, and S. Lloyd, “Efficient classical algorithms for simulating symmetric quantum systems,” *Quantum* **7**, 1189 (2023).
- [36] M. L. Goh, M. Larocca, L. Cincio, M. Cerezo, and F. Sauvage, “Lie-algebraic classical simulations for quantum computing,” *Phys. Rev. Res.* **7**, 033266 (2025).
- [37] M. Cerezo, M. Larocca, D. García-Martín, N. L. Diaz, P. Braccia, E. Fontana, M. S. Rudolph, P. Bermejo, A. Ijaz, S. Thanasilp, E. R. Anschuetz, and Z. Holmes, “Does provable absence of barren plateaus imply classical simulability?” *Nat. Commun.* **16**, 7907 (2025).
- [38] X. Gao, E. R. Anschuetz, S.-T. Wang, J. I. Cirac, and M. D. Lukin, “Enhancing generative models via quantum correlations,” *Phys. Rev. X* **12**, 021037 (2022).
- [39] H. Zhao and D.-L. Deng, “Entanglement-induced provable and robust quantum learning advantages,” (2024), [arXiv:2410.03094 \[quant-ph\]](#).
- [40] E. R. Anschuetz, H.-Y. Hu, J.-L. Huang, and X. Gao, “Interpretable quantum advantage in neural sequence learning,” *PRX Quantum* **4**, 020338 (2023).

- [41] R. Babbush, J. R. McClean, M. Newman, C. Gidney, S. Boixo, and H. Neven, “Focus beyond quadratic speedups for error-corrected quantum advantage,” *PRX Quantum* **2**, 010103 (2021).
- [42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, USA, 2017) pp. 6000–6010.
- [43] Y. Takeuchi, A. Mantri, T. Morimae, A. Mizutani, and J. F. Fitzsimons, “Resource-efficient verification of quantum computing using serfling’s bound,” *npj Quantum Inf.* **5**, 27 (2019).
- [44] M. A. Webster, B. J. Brown, and S. D. Bartlett, “The XP stabiliser formalism: a generalisation of the Pauli stabiliser formalism with arbitrary phases,” *Quantum* **6**, 815 (2022).
- [45] R. Somma, H. Barnum, G. Ortiz, and E. Knill, “Efficient solvability of Hamiltonians and limits on the power of some quantum computational models,” *Phys. Rev. Lett.* **97**, 190501 (2006).
- [46] M. Mints and E. R. Anschuetz, “Fragmentation is efficiently learnable by quantum neural networks,” (2025), [arXiv:2512.00751 \[quant-ph\]](https://arxiv.org/abs/2512.00751).
- [47] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. Aga, J. Huang, C. Bai, M. Gschwind, A. Gupta, M. Ott, A. Melnikov, S. Candido, D. Brooks, G. Chauhan, B. Lee, H.-H. Lee, B. Akyildiz, M. Balandat, J. Spisak, R. Jain, M. Rabbat, and K. Hazelwood, “Sustainable AI: Environmental implications, challenges and opportunities,” in *Proceedings of Machine Learning and Systems*, Vol. 4, edited by D. Marculescu, Y. Chi, and C. Wu (2022) pp. 795–813.
- [48] I. Mavromatis, K. Katsaros, and A. Khan, “Computing within limits: An empirical study of energy consumption in ML training and inference,” (2024), [arXiv:2406.14328 \[cs.LG\]](https://arxiv.org/abs/2406.14328).
- [49] S. Kochen and E. P. Specker, “The problem of hidden variables in quantum mechanics,” *J. Math. Mech.* **17**, 59 (1967).
- [50] S. Abramsky and A. Brandenburger, “The sheaf-theoretic structure of non-locality and contextuality,” *New J. Phys.* **13**, 113036 (2011).
- [51] S. Abramsky and M. Sadrzadeh, “Semantic unification,” in *Categories and Types in Logic, Language, and Physics: Essays Dedicated to Jim Lambek on the Occasion of His 90th Birthday*, edited by C. Casadio, B. Coecke, M. Moortgat, and P. Scott (Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, 2014) pp. 1–13.
- [52] S. Abramsky, “Contextual semantics: From quantum mechanics to logic, databases, constraints, and complexity,” in *Contextuality from Quantum Physics to Psychology* (World Scientific Publishing, 2016) Chap. Chapter 2, pp. 23–50.
- [53] S. Abramsky, R. Soares Barbosa, K. Kishida, R. Lal, and S. Mansfield, “Contextuality, cohomology and paradox,” in *24th EACSL Annual Conference on Computer Science Logic (CSL 2015)*, Leibniz International Proceedings in Informatics (LIPIcs), Vol. 41, edited by S. Kreutzer (Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2015) pp. 211–228.
- [54] V. H. Cervantes and E. N. Dzhafarov, “Contextuality analysis of impossible figures,” *Entropy* **22**, 981 (2020).
- [55] R. Raz, “Exponential separation of quantum and classical communication complexity,” in *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, STOC ’99 (Association for Computing Machinery, New York, NY, USA, 1999) pp. 358–367.
- [56] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf, “Quantum fingerprinting,” *Phys. Rev. Lett.* **87**, 167902 (2001).
- [57] F. Le Gall, “Exponential separation of quantum and classical online space complexity,” in *Proceedings of the Eighteenth Annual ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA ’06, edited by P. B. Gibbons and U. Vishkin (Association for Computing Machinery, New York, NY, USA, 2006) pp. 67–73.
- [58] D. Gavinsky, J. Kempe, I. Kerenidis, R. Raz, and R. de Wolf, “Exponential separations for one-way quantum communication complexity, with applications to cryptography,” in *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, STOC ’07, edited by

- D. Johnson and U. Feige (Association for Computing Machinery, New York, NY, USA, 2007) pp. 516–525.
- [59] S. Gupta, D. Saha, Z.-P. Xu, A. Cabello, and A. S. Majumdar, “Quantum contextuality provides communication complexity advantage,” *Phys. Rev. Lett.* **130**, 080802 (2023).
 - [60] J. Kallaugher, O. Parekh, and N. Voronova, “Exponential quantum space advantage for approximating maximum directed cut in the streaming model,” (2023), [arXiv:2311.14123 \[quant-ph\]](#).
 - [61] S. Manna, A. Chaturvedi, and D. Saha, “Unbounded quantum advantage in communication complexity measured by distinguishability,” (2024), [arXiv:2401.12903 \[quant-ph\]](#).
 - [62] M. S. Rudolph, S. Lerch, S. Thanasilp, O. Kiss, S. Vallecorsa, M. Grossi, and Z. Holmes, “Trainability barriers and opportunities in quantum generative modeling,” (2023), [arXiv:2305.02881 \[quant-ph\]](#).
 - [63] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, edited by Z. Ghahramani, M. Welling, C. Cortes, and N. Lawrence (Curran Associates, Inc., Red Hook, NY, USA, 2014) pp. 3104–3112.
 - [64] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, “A comparison of sequence-to-sequence models for speech recognition,” in *Proc. Interspeech 2017*, edited by F. Lacerda (Curran Associates, Inc., Red Hook, NY, USA, 2017) pp. 939–943.
 - [65] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by K. Barnard, H. Bischof, P. Felzenszwalb, D. Forsyth, S. Lazebnik, and J. Matas (Curran Associates, Inc., Red Hook, NY, USA, 2015) pp. 3156–3164.
 - [66] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554 (1982).
 - [67] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.* **9**, 1735 (1997).
 - [68] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by A. Moschitti, B. Pang, and W. Daelemans (Association for Computational Linguistics, Doha, 2014) pp. 1724–1734.
 - [69] A. Mokhtari, A. Ozdaglar, and A. Jadbabaie, “Efficient nonconvex empirical risk minimization via adaptive sample size methods,” in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 89, edited by K. Chaudhuri and M. Sugiyama (PMLR, Naha, Japan, 2019) pp. 2485–2494.
 - [70] S. Mei, Y. Bai, and A. Montanari, “The landscape of empirical risk for nonconvex losses,” *Ann. Stat.* **46**, 2747 (2018).
 - [71] J. Yang, W. Hu, and C. J. Li, “On the fast convergence of random perturbations of the gradient flow,” *Asymptot. Anal.* **122**, 371 (2021).
 - [72] R. Dixit, M. Gurbuzbalaban, and W. U. Bajwa, “Accelerated gradient methods for nonconvex optimization: Escape trajectories from strict saddle points and convergence to local minima,” (2023), [arXiv:2307.07030 \[math.OC\]](#).
 - [73] S. V. Kurochkin, “Neural network with smooth activation functions and without bottlenecks is almost surely a Morse function,” *Comput. Math. and Math. Phys.* **61**, 1162 (2021).
 - [74] N. Liu, J. Thompson, C. Weedbrook, S. Lloyd, V. Vedral, M. Gu, and K. Modi, “Power of one qumode for quantum computation,” *Phys. Rev. A* **93**, 052304 (2016).
 - [75] D. Gottesman, A. Kitaev, and J. Preskill, “Encoding a qubit in an oscillator,” *Phys. Rev. A* **64**, 012310 (2001).
 - [76] M. A. Nielsen and I. L. Chuang, “The quantum Fourier transform and its applications,” in *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, UK, 2010) pp. 216–247.
 - [77] M. J. Bremner, A. Montanaro, and D. J. Shepherd, “Achieving quantum supremacy with sparse and noisy commuting quantum computations,” *Quantum* **1**, 8 (2017).

- [78] N. D. Mermin, “Simple unified form for the major no-hidden-variables theorems,” *Phys. Rev. Lett.* **65**, 3373 (1990).
- [79] D. Gottesman and I. L. Chuang, “Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations,” *Nature* **402**, 390 (1999).
- [80] P. J. Rabier, “Ehresmann fibrations and Palais-Smale conditions for morphisms of Finsler manifolds,” *Ann. Math.* **146**, 647 (1997).
- [81] L. Isenhower, M. Saffman, and K. Mølmer, “Multibit C_k NOT quantum gates via Rydberg blockade,” *Quantum. Inf. Process.* **10**, 755 (2011).
- [82] D. Bluvstein, S. J. Evered, A. A. Geim, S. H. Li, H. Zhou, T. Manovitz, S. Ebadi, M. Cain, M. Kalinowski, D. Hangleiter, *et al.*, “Logical quantum processor based on reconfigurable atom arrays,” *Nature*, 1 (2023).
- [83] A. J. Kerman, “Quantum information processing using quasiclassical electromagnetic interactions between qubits and electrical resonators,” *New J. Phys.* **15**, 123011 (2013).
- [84] A. Blais, A. L. Grimsmo, S. M. Girvin, and A. Wallraff, “Circuit quantum electrodynamics,” *Rev. Mod. Phys.* **93**, 025005 (2021).
- [85] M. Leifer, “Is the quantum state real? an extended review of ψ -ontology theorems,” *Quanta* **3**, 67 (2014).
- [86] S. Bravyi, D. Gosset, R. Koenig, and M. Tomamichel, “Quantum advantage with noisy shallow circuits,” *Nat. Phys.* **16**, 1040 (2020).
- [87] L. Caha, X. Coiteux-Roy, and R. Koenig, “A colossal advantage: 3D-local noisy shallow quantum circuits defeat unbounded fan-in classical circuits,” (2023), [arXiv:2312.09209 \[quant-ph\]](#).
- [88] M. H. Teo, W. Yang, J. Sud, T. Tomesh, F. T. Chong, and E. R. Anschuetz, “k-contextuality as a heuristic for memory separations in learning,” (2025), [arXiv:2507.11604 \[quant-ph\]](#).
- [89] T. Brown *et al.*, “Language models are few-shot learners,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, NY, USA, 2020) pp. 1877–1901.
- [90] OpenAI, “GPT-4 technical report,” (2023), [arXiv:2303.08774 \[cs.CL\]](#).
- [91] K. Wan, S. Choi, I. H. Kim, N. Shutty, and P. Hayden, “Fault-tolerant qubit from a constant number of components,” *PRX Quantum* **2**, 040345 (2021).
- [92] M. Srikumar, S. D. Bartlett, and A. Karanjai, “How contextuality and antidistinguishability are related,” (2024), [arXiv:2411.09919 \[quant-ph\]](#).
- [93] A. Mari and J. Eisert, “Positive Wigner functions render classical simulation of quantum computation efficient,” *Phys. Rev. Lett.* **109**, 230503 (2012).
- [94] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Curran Associates, Inc., Red Hook, NY, USA, 2014) pp. 2672–2680.
- [95] B. Shi, W. J. Su, and M. I. Jordan, “On learning rates and Schrödinger operators,” (2020), [arXiv:2004.06977 \[cs.LG\]](#).
- [96] R. Duan, H. Wu, and R. Zhou, “Faster matrix multiplication via asymmetric hashing,” (2022), [arXiv:2210.10173 \[cs.DS\]](#).
- [97] V. Strassen, “Gaussian elimination is not optimal,” *Numer. Math.* **13**, 354 (1969).
- [98] S. Aaronson and D. Gottesman, “Improved simulation of stabilizer circuits,” *Phys. Rev. A* **70**, 052328 (2004).
- [99] C. Calcluth, A. Ferraro, and G. Ferrini, “Efficient simulation of Gottesman-Kitaev-Preskill states with Gaussian circuits,” (2022), [arXiv:2203.11182 \[quant-ph\]](#).
- [100] M. Larocca, P. Czarnik, K. Sharma, G. Muraleedharan, P. J. Coles, and M. Cerezo, “Diagnosing barren plateaus with tools from quantum optimal control,” *Quantum* **6**, 824 (2022).

A Weighted Hypergraph States

A.1 Qubit Weighted Hypergraph States

We here give a bevy of facts about *qubit weighted hypergraph states*. Though first introduced in Ref. [44] in the context of the so-called XP stabilizer formalism, we here directly consider a stabilizer formalism of such states.

We first review the definition of qubit weighted hypergraph states, focusing for simplicity on cases where the associated hypergraphs are *k-uniform hypergraphs*; namely, their only nontrivial hyperedges are associated with cardinality k subsets of vertices. Let G be a k -uniform weighted hypergraph with hyperedges labeled by sets $\bar{v}$ and with associated hyperedge weights $e_{\bar{v}}$. We denote as $|G\rangle$ the *qubit k-uniform weighted hypergraph state associated with G*:

$$|G\rangle := \prod_{\bar{v} \in G} U_{\bar{v}}(e_{\bar{v}}) |+\rangle^{\otimes n}, \quad (29)$$

where we use the shorthand:

$$U_{\bar{v}}(e_{\bar{v}}) := C^{k-1} Z_{\bar{v}}(-e_{\bar{v}}) = \exp\left(i\pi e_{\bar{v}} \prod_{i \in \bar{v}} \frac{I - Z_i}{2}\right) \quad (30)$$

to ease on notation. It is easy to check the commutation relations when $i \in \bar{v}$:

$$X_i U_{\bar{v}}(e_{\bar{v}}) = U_{\bar{v} \setminus \{i\}}(e_{\bar{v}}) U_{\bar{v}}(-e_{\bar{v}}) X_i, \quad (31)$$

$$U_{\bar{v}}(e_{\bar{v}}) X_i = X_i U_{\bar{v} \setminus \{i\}}(e_{\bar{v}}) U_{\bar{v}}(-e_{\bar{v}}). \quad (32)$$

By construction, $|G\rangle$ is the mutual +1 eigenstate of the commuting operators:

$$\begin{aligned} s_i &= \prod_{\bar{v} \ni i} U_{\bar{v}}(e_{\bar{v}}) X_i \left(\prod_{\bar{v} \ni i} U_{\bar{v}}(e_{\bar{v}}) \right)^\dagger \\ &= X_i \prod_{\bar{v} \ni i} U_{\bar{v} \setminus \{i\}}(e_{\bar{v}}) U_{\bar{v}}(-2e_{\bar{v}}) \\ &= \left(\prod_{\bar{v} \ni i} U_{\bar{v} \setminus \{i\}}(-e_{\bar{v}}) U_{\bar{v}}(2e_{\bar{v}}) \right) X_i. \end{aligned} \quad (33)$$

From this it is immediately clear that the s_i are their own inverses:

$$s_i^2 = I. \quad (34)$$

Thus,

$$s_i = \exp\left(-i\frac{\pi}{2}\right) \exp\left(i\frac{\pi}{2}s_i\right) \quad (35)$$

and in particular $|G\rangle$ is also stabilized by all powers s_i^α . It is also apparent from Eq. (33) that the s_i are Hermitian.

We now consider the special case where G has only a single hyperedge $\bar{v}$. Consider the stabilizer of $|G\rangle$:

$$s_{\bar{v}} := \prod_{i \in \bar{v}} s_i \quad (36)$$

associated with some vertex set $\bar{v}$. By Eq. (32) it is apparent this can be written as:

$$s_{\bar{v}} = \left(\prod_{i \in \bar{v}} X_i \right) \exp(i\pi e_{\bar{v}} R), \quad (37)$$

where R is some $e_{\bar{v}}$ -independent multilinear degree- k polynomial in the $\frac{I-Z_i}{2}$ with integer coefficients. As the commutation in Eq. (32) only introduces terms of up to a single degree lower than the commuted $U_{\bar{v}}$, we can also calculate the constant term R_0 in R :

$$R_0 = 1. \quad (38)$$

In particular,

$$\exp(i\pi e_{\bar{v}} R) |0\rangle^{\otimes n} = \exp(i\pi e_{\bar{v}}) |0\rangle^{\otimes n}. \quad (39)$$

Similarly, by Eq. (31)

$$s_{\bar{v}} = \exp(i\pi e_{\bar{v}} L) \prod_{i \in \bar{v}} X_i, \quad (40)$$

where once again L is some $e_{\bar{v}}$ -independent multilinear degree- k polynomial in the $\frac{I-Z_i}{2}$. By the Hermiticity of $s_{\bar{v}}$ we must have that:

$$L = -R. \quad (41)$$

In particular, when $e_{\bar{v}}$ is an integer, we have by equating these expressions for $s_{\bar{v}}$ that:

$$\left[\cos(\pi e_{\bar{v}} (R - 1)), \prod_{i \in \bar{v}} X_i \right] = \mathbf{0}, \quad (42)$$

where:

$$\cos(\pi e_{\bar{v}} (R - 1)) |0\rangle^{\otimes n} = |0\rangle^{\otimes n}. \quad (43)$$

A.2 Qumode Weighted Hypergraph States

One can consider a qumode version of k -uniform weighted hypergraph states completely analogously to qubit weighted hypergraph states [43]. The standard definition of such states involves the sequential application of qumode multi-controlled phase operators associated with a k -uniform hypergraph G on a momentum-squeezed state:

$$|G\rangle := \prod_{\bar{v} \in G} \exp\left(i\pi e_{\bar{v}} \prod_{j \in \bar{v}} \hat{q}_j\right) |\mathbf{0}\rangle_{\hat{p}}. \quad (44)$$

Here, we use the notation $|\mathbf{a}\rangle_{\hat{p}} = \bigotimes |a_i\rangle_{\hat{p}_i} = \bigotimes |\hat{p}_i = a_i\rangle$. This is the definition presented in Sec. 3.1.

In the formal proof of our results—presented in Appendix B—we will consider a slightly modified definition. Rather than applying gates to a squeezed momentum state, we consider an initial *GKP state* [75]:

$$|\text{GKP}\rangle \propto \bigotimes_{i=1}^n \left(\sum_{\alpha=-\infty}^{\infty} |\alpha\rangle_{\hat{q}_i} \right). \quad (45)$$

Specifically, we take *qumode k -uniform weighted hypergraph states* associated with a k -uniform hypergraph G to be defined as:

$$|G\rangle := \prod_{\bar{v} \in G} U_{\bar{v}}(e_{\bar{v}}) |\text{GKP}\rangle, \quad (46)$$

where now in a qumode context we have defined:

$$U_{\bar{v}}(e_{\bar{v}}) := \exp\left(i\pi e_{\bar{v}} \prod_{j \in \bar{v}} \hat{q}_j\right). \quad (47)$$

It is once again easy to check the commutation relations when $i \in \bar{v}$:

$$X_i U_{\bar{v}}(e_{\bar{v}}) = U_{\bar{v} \setminus \{i\}}(-e_{\bar{v}}) U_{\bar{v}}(e_{\bar{v}}) X_i, \quad (48)$$

$$U_{\bar{v}}(e_{\bar{v}}) X_i = X_i U_{\bar{v} \setminus \{i\}}(e_{\bar{v}}) U_{\bar{v}}(e_{\bar{v}}), \quad (49)$$

where now:

$$X_i := \exp(-2i\hat{p}_i). \quad (50)$$

Similarly to qubit k -uniform weighted hypergraph states, the qumode version of these states are described by commuting stabilizers:

$$\begin{aligned} s_i &= \prod_{\bar{v} \ni i} U_{\bar{v}}(e_{\bar{v}}) X_i \left(\prod_{\bar{v} \ni i} U_{\bar{v}}(e_{\bar{v}}) \right)^\dagger \\ &= X_i \prod_{\bar{v} \ni i} U_{\bar{v} \setminus \{i\}}(e_{\bar{v}}) \end{aligned} \quad (51)$$

and their Hermitian conjugates. Unlike the qubit case, however, the s_i are now not Hermitian.

We now consider the special case where G has only a single hyperedge $\bar{v}$. Consider the stabilizer of $|G\rangle$:

$$s_{\bar{v}} := \prod_{i \in \bar{v}} s_i \quad (52)$$

associated with some vertex set $\bar{v}$. By Eq. (49) it is apparent this can be written as:

$$s_{\bar{v}} = \left(\prod_{i \in \bar{v}} X_i \right) \exp(i\pi e_{\bar{v}} R), \quad (53)$$

where R is some $e_{\bar{v}}$ -independent multilinear degree- $(k-1)$ polynomial in the $\hat{q}_i$ with integer coefficients. As the commutation in Eq. (49) only introduces terms of up to a single degree lower than the commuted $U_{\bar{v}}$, we can also calculate the constant term R_0 in R :

$$R_0 = 1. \quad (54)$$

In particular,

$$\exp(i\pi e_{\bar{v}} R) |\mathbf{0}\rangle_{\hat{q}} = \exp(i\pi e_{\bar{v}}) |\mathbf{0}\rangle_{\hat{q}}. \quad (55)$$

Unlike the qubit case, the $s_{\bar{v}}$ are now *not* Hermitian, and particular it is not true that $\exp(i\pi e_{\bar{v}} R)$ commutes through $\prod_{i \in \bar{v}} X_i$ up to conjugation. However, this *is* true up to stabilizers of $|G\rangle$ when $e_{\bar{v}}$ is an integer. Namely, using Eq. (49) we have that:

$$\exp(i\pi e_{\bar{v}} R) \left(\prod_{i \in \bar{v}} X_i \right) = \left(\prod_{i \in \bar{v}} X_i \right) \exp(i\pi e_{\bar{v}} R) \exp(2i\pi e_{\bar{v}} \tilde{R}), \quad (56)$$

where $\tilde{R}$ is another $e_{\bar{v}}$ -independent multilinear degree- k polynomial in the $\hat{q}_i$ with integer coefficients and no constant term. As $\exp(2i\pi e_{\bar{v}} \tilde{R})$ commutes with $U_{\bar{v}}(e_{\bar{v}})$ and as $|G\rangle$ is a uniform superposition state over squeezed integer positions,

$$\exp(i\pi e_{\bar{v}} R) \left(\prod_{i \in \bar{v}} X_i \right) |G\rangle = \left(\prod_{i \in \bar{v}} X_i \right) \exp(i\pi e_{\bar{v}} R) |G\rangle. \quad (57)$$

This is similar to an idea used in the work of Ref. [91], where generated high-weight errors in a proposed fault-tolerant quantum computing scheme were shown to be equivalent to low-weight errors up to stabilizers. We thus have when $e_{\bar{v}}$ is an integer that for all integer α :

$$\left[\exp(i\alpha\pi e_{\bar{v}} (R-1)), \prod_{i \in \bar{v}} X_i \right] |G\rangle = \mathbf{0}, \quad (58)$$

where:

$$\exp(i\pi e_{\bar{v}} (R-1)) |\mathbf{0}\rangle_{\hat{q}} = |\mathbf{0}\rangle_{\hat{q}}. \quad (59)$$

As the projector onto the $\exp(i\pi e_{\bar{v}} (R-1)) = 1$ subspace can be written as integer powers of $\exp(i\pi e_{\bar{v}} (R-1))$, this implies that the measurement (given this measurement outcome) commutes with the product of vertex stabilizers of $|G\rangle$.

B Proofs of Expressivity Separations

B.1 (ℓ, n, k) -Hypergraph Stabilizer Measurement Translation

Before giving proofs of expressivity separations between our quantum model and classical models, we first give a formal definition of the translation task we will prove a separation on: namely, (ℓ, n, k) -hypergraph stabilizer measurement translation ((ℓ, n, k) -HSMT), parameterized by ℓ , n , and k , where $\ell \geq \binom{n}{k} + 3n - k$. Note that the technical description of the task described here differs from construction presented in the main text, mainly in that there are additional constraints on the forms the various input tokens $\mathbf{x}_i$ can take. The fact that the k -HRNN described in Sec. 3 (or a slight variant thereof) can perfectly perform the task using $O(n)$ qubits or qumodes will follow immediately. In what follows, we use the notation $\binom{[n]}{r}$ to denote the set of r -tuples with distinct elements all drawn from $\{1, \dots, n\}$.

We consider an input language given by ℓ -long sequences of tokens:

$$\mathbf{x} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\ell)})^\top. \quad (60)$$

Recall from the main text that a qubit k -HRNN has initial latent state $|\lambda_1\rangle = |0\rangle^{\otimes n}$, and a qumode k -HRNN has initial latent state $|\lambda_1\rangle = |0\rangle_{\hat{q}}$ (i.e., it is initially squeezed in the position basis). Just as in the main text, tokens in the (ℓ, n, k) -HSMT task will be of the form $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\phi}, \boldsymbol{\theta})$, describing parameter settings of a k -HRNN. However, to aid in our proofs, we assume a variant of the translation task with certain additional constraints on the tokens.

We assume that the first n tokens are of the form:

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\phi}, \boldsymbol{\theta}^{(i)}) = (0, i, 0, -2v_i, 0); \quad (61)$$

i.e., it corresponds to a k -HRNN measuring either $\exp(2iv_i X_i)$ (in the qubit setting) or $\exp(2iv_i \hat{p}_i)$ (in the qumode setting) via phase estimation. In the qumode setting, when $v_i = 1$ and the measurement outcomes are $+1$, this sequence of measurements projects onto the GKP state described in Eq. (45); this can be seen by considering the phase estimation circuit [74] measuring $O_{\theta_i, p}$ in the position basis, and considering its action on the state $|0\rangle_{\hat{q}_i}$.

The next $\binom{n}{k}$ tokens—labeled by k -tuples $\bar{v} \in \binom{[n]}{k}$ in an arbitrary order—are of the form:

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\phi}, \boldsymbol{\theta}^{(i)}) = (\bar{v}, 0, -\gamma_{\bar{v}}, 0, 0), \quad (62)$$

i.e., it corresponds to a k -HRNN applying $C^{k-1} Z_{\bar{v}}(\gamma_{\bar{v}})$ sequentially.

The next n tokens query the model to reproduce the first n measurement results (or are excluded, if we are studying hypergraph stabilizer measurement *without repetition*). This is essentially just a technical assumption that will allow us to assume any states indistinguishable by a classical model at this point in the sequence have identical outputs. This can be captured by a slightly-modified k -HRNN architecture which explicitly stores its first n measurement results in an extra n qumodes of memory.

The next $n - k$ tokens are of the form:

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\phi}, \boldsymbol{\theta}) = (0, b_i, 0, 0, \theta_{\{b_i\}} = \epsilon) \quad (63)$$

for distinct integer b_i and $\epsilon = 1$ (in the qubit setting) or $\epsilon \rightarrow 0$ (in the qumode setting). That is, it corresponds to a k -HRNN measuring Z_{b_i} (in the qubit setting) or $\hat{q}_{b_i}$ (in the qumode setting) via phase estimation.¹

Finally, all remaining tokens are of the form:

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\phi}, \boldsymbol{\theta}) = (0, \bar{c}, 0, -\boldsymbol{\phi}, -\boldsymbol{\theta}) \quad (64)$$

¹In practice, one would change the ancillary register $|a\rangle$ from $|GKP\rangle$ to $|0\rangle_{\hat{p}}$ in performing phase estimation at this stage in the qumode setting; this only slightly modifies the k -HRNN architecture.

where $\bar{c}$ is the remaining k qumodes that were not one of the b_i , i.e., it corresponds to a general k -HRNN measuring general G . on the k qumodes previously unmeasured.

By construction, all input sequences $\mathbf{x}$ are then described by $\mathbf{v} \in \mathbb{R}^n$, $\boldsymbol{\gamma} \in \mathbb{R}^{\binom{n}{k}}$, $\bar{b} \in \binom{[n]}{n-k}$, and (ϕ, θ) describe the remainder of the sequence. A translation $\mathbf{y}$ of $\mathbf{x}$ is considered correct if it is of the form:

$$\mathbf{y} = (m_1, \dots, m_n, 0, \dots, 0, m_1, \dots, m_n, m_{\ell-3n+1}, \dots, m_\ell)^\top, \quad (65)$$

where the measurement outcomes m_i are consistent with those of a k -HRNN performing the previously described procedure.

B.2 Single-copy Antidistinguishing Measurement Sequences via Quantum Contextuality

We now prove two lemmas demonstrating the existence of *antidistinguishing measurement sequences* given by qubit k -uniform weighted hypergraph state stabilizers, as well as qumode k -uniform hypergraph state stabilizers. Antidistinguishing measurement sequences are those which, given r candidates for a given quantum state, rule out one candidate with certainty [85]. Their existence is equivalent to the presence of quantum contextuality in a set of quantum states [92].

We first prove that antidistinguishing measurement sequences via the measurement of quantum contextual observables between qubit k -uniform hypergraph states exist.

Lemma B.1 (Qubit k -uniform hypergraph states are antidistinguishable). *Consider the state $|0\rangle^{\otimes n}$. Let $|\psi_1\rangle$ and $|\psi_2\rangle$ be two qubit k -uniform weighted hypergraph states with some hyperedge weight differing by 1. There exists an antidistinguishing measurement sequence (via measuring quantum contextual observables of the form of G .) of length $n - k + 2$ that with certainty discounts one of the three states.*

Proof. Let $\bar{v}$ be the set of vertices associated with a hyperedge where $|\psi_1\rangle$ and $|\psi_2\rangle$ have weights $e'_{\bar{v}}, e''_{\bar{v}}$, respectively, differing by 1. For simplicity, we consider the basis transformed by powers of $C^k Z$ operations such that $e'_{\bar{v}} = 0$ and $e''_{\bar{v}} = 1$; note that this can be done without loss of generality (WLOG) as this basis transformation maps the class of operators comprising the qubit k -HRNN to itself, as well as maps $|0\rangle^{\otimes n}$ to itself.

The antidistinguishing measurement sequence first consists of measuring Z_i for all qubits labeled by i associated with the $n - k$ vertices not in $\bar{v}$; if one of these measurement results is not 1, then $|0\rangle^{\otimes n}$ is discounted (i.e., the state in question is known not to be $|0\rangle^{\otimes n}$). We assume then that they are all 1. For all $i \in \bar{v}$, we have by definition that the post-measurement state of $|\psi_1\rangle$ is then just:

$$|+\rangle^{\otimes n}, \quad (66)$$

and the post-measurement state of $|\psi_2\rangle$ just the unweighted qubit graph state with stabilizers:

$$s_i = X_i C^k Z_{\bar{v} \setminus \{i\}}. \quad (67)$$

Consider now the measurement of the operator:

$$M = \cos(\pi(R - 1)), \quad (68)$$

where R is defined as in Eq. (37). As $\exp(i\pi(R - 1))$ is a product of multi-controlled Z operators, and as multi-controlled Z operators are Hermitian, M can be written as a product of multi-controlled Z operators.

By the discussion surrounding Eq. (37), $R - 1$ nullifies $|0\rangle^{\otimes n}$. Thus, the measurement result of M must be $+1$; otherwise, $|0\rangle^{\otimes n}$ would be distinguished. Furthermore, from Eq. (42), this measurement commutes with the product of vertex stabilizers for $|\psi_1\rangle, |\psi_2\rangle$. Note also that the projection onto the subspace where $M = 1$ is equal to the projection onto the subspace where $\exp(i\pi(R - 1)) = 1$. Thus,

the post-measurement states $|\phi_1\rangle, |\phi_2\rangle$ of $|\psi_1\rangle, |\psi_2\rangle$, respectively, then are respectively stabilized by:

$$t' = \prod_{i \in \bar{v}} X_i, \quad (69)$$

$$t'' = - \prod_{i \in \bar{v}} X_i. \quad (70)$$

$|\phi_1\rangle$ and $|\phi_2\rangle$ are thus orthogonal with distinguishing measurement given by t' . $\square$

The antidistinguishing measurement sequence for qumodes is similar.

Lemma B.2 (Qumode k -uniform hypergraph states are antidistinguishable). *Consider the state $|\mathbf{0}\rangle_{\hat{q}}$. Let $|\psi_1\rangle$ and $|\psi_2\rangle$ be two qumode k -uniform weighted hypergraph states with some hyperedge weight differing by 1. There exists an antidistinguishing measurement sequence (via measuring quantum contextual observables of the form of G .) of length $n - k + 2$ that with certainty discounts one of the three states.*

Proof. Let $\bar{v}$ be the set of vertices associated with a hyperedge where $|\psi_1\rangle$ and $|\psi_2\rangle$ have weights $e'_{\bar{v}}, e''_{\bar{v}}$, respectively, differing by 1. We assume WLOG that $e''_{\bar{v}} < e'_{\bar{v}}$; note that we cannot assume WLOG anymore that one is 0 and the other 1, as the class of operators $\exp(2i\phi\hat{p}_i)$ no longer maps to itself under conjugation by the required basis transformation.

The antidistinguishing measurement sequence first consists of measuring $\hat{q}_i$ for all qubits labeled by i associated with the $n - k$ vertices not in $\bar{v}$; if one of these measurement results is not 0, then $|\mathbf{0}\rangle_{\hat{q}}$ is discounted (i.e., the state in question is known not to be $|\mathbf{0}\rangle_{\hat{q}}$). We assume then that they are all 0. For all $i \in \bar{v}$, we have by definition that the post-measurement state of $|\psi_1\rangle$ is then just the k -uniform weighted qumode graph state with stabilizers::

$$s'_i = X_i \exp\left(i\pi e'_{\bar{v}} \prod_{j \in \bar{v}} \hat{q}_j\right), \quad (71)$$

and the post-measurement state of $|\psi_2\rangle$ just the unweighted qumode graph state with stabilizers:

$$s''_i = X_i \exp\left(i\pi e''_{\bar{v}} \prod_{j \in \bar{v}} \hat{q}_j\right). \quad (72)$$

Consider now the measurement of the operator:

$$M = \exp(i\pi(R - 1)), \quad (73)$$

where R is defined as in Eq. (53).

By the discussion surrounding Eq. (37), $\exp(i\pi(R - 1))$ stabilizes $|\mathbf{0}\rangle_{\hat{q}}$. Thus, the measurement result of M must be +1; otherwise, $|\mathbf{0}\rangle_{\hat{q}}$ would be distinguished. Furthermore, from Eq. (58), this measurement commutes with the product of vertex stabilizers for $|\psi_1\rangle, |\psi_2\rangle$, up to stabilizers of $|\psi_1\rangle, |\psi_2\rangle$. Thus, the post-measurement states $|\phi_1\rangle, |\phi_2\rangle$ of $|\psi_1\rangle, |\psi_2\rangle$, respectively, then are respectively stabilized by:

$$t' = s', \quad (74)$$

$$t'' = \exp(i\pi(e''_{\bar{v}} - e'_{\bar{v}})) s' = -s'. \quad (75)$$

$|\phi_1\rangle$ and $|\phi_2\rangle$ are thus orthogonal with distinguishing measurement given by t' . $\square$

B.3 Expressivity Separation for Autoregressive Models

We consider now an autoregressive learner with structure given by Fig. 1(a). We assume for simplicity that the learner is deterministic after given a random vector $\mathbf{r}$. This class of models includes implementations of stochastic simulation algorithms such as Wigner function simulation [93] as well as generative adversarial networks (GANs) [94]. As for each $\mathbf{r}$, we will demonstrate that there exists an input sequence such that any classical model with $\dim(L) < \binom{n}{k} - 1$ deterministically outputs a measurement sequence inconsistent with quantum mechanics, our results will still hold when considering the model over its randomness. Due to this we will often take the $\mathbf{r}$ -dependence to be implicit.

Autoregressive neural sequence models at time step i map a d -dimensional input token $\mathbf{x}_i$ and a latent vector $\boldsymbol{\lambda}_{i-1}$ to an output token y_i and a new latent vector $\boldsymbol{\lambda}_i$. After ℓ steps, then, we can consider the autoregressive model as a function:

$$\mathcal{F}^{\mathbf{r}} : (\mathbb{R}^d)^\ell \rightarrow L \times \mathbb{R}^\ell \quad (76)$$

where $\mathbf{r}$ is a random vector such that, for any $\mathbf{r}$, $\mathcal{F}^{\mathbf{r}}$ is deterministic. In the following we will assume that L is a C^2 contractible Finsler manifold; concrete examples of such a manifold are the open ball $L = (0, 1)^d$ and the reals $L = \mathbb{R}^D$ for some D .

We now state our conditions on $\mathcal{F}^{\mathbf{r}}$, stated informally in the main text as Assumptions 2.1 and 2.2. Our proof of Theorem B.3 concern a subspace K of inputs corresponding to measurements of hypergraph state stabilizers. We require that the projection of $\mathcal{F}^{\mathbf{r}}$ onto L is C^2 on a contractible, $(\dim(L) + 2)$ -dimensional, open subset $U \subseteq K$. As $\dim(K) = \binom{n}{k}$, our assumption on the dimension of U is satisfied when U is not measure zero in K and $L < \binom{n}{k} - 1$, as stated in the main text. We also require that on U , $\mathcal{F}^{\mathbf{r}}$ is a *strong submersion* [80]—namely, U is such that the Jacobian of $\mathcal{F}^{\mathbf{r}}$ has a $\dim(L) \times \dim(L)$ minor with lower-bounded determinant. This is effectively a requirement that the model is *strongly Morse* [69] in a multivariate sense. Neural networks being strongly Morse (in the univariate case) is a common assumption [69–72], and certain classes of neural networks are known to be almost surely Morse over settings of their parameters [73]. Finally, we require as a technical assumption that no sequence with limit point in $\partial U \not\subseteq U$ maps to a sequence with limit point in L under $\mathcal{F}^{\mathbf{r}}$ (projected onto L). By the “global implicit function theorem” of Ref. [80], $\mathcal{F}^{\mathbf{r}}$ (projected onto L) is then “globally” (on U) a trivial bundle with base space C^1 -diffeomorphic to L . This, along with Lemmas B.1 and B.2, are the central implications which we will use to prove our theorems.

We give three simple examples that immediately imply our strong submersion condition on (an assumed C^2) $\mathcal{F}^{\mathbf{r}}$:

1. The presence of an appropriate ℓ_2 regularization term dominating at infinity.
2. More generally, assuming that updates to the model during training (via, for instance, gradient descent) are bounded, as our conditions are implied by regularity of the model at infinity. A similar approach was taken in the work of Ref. [95] in justifying certain regularity conditions of machine learning models at infinity.
3. The model is implemented at a finite precision $\epsilon \ll 1$. A critical value, by definition, has at least one local coordinate that varies only by $\sim \epsilon^2 \approx 0$ for an ϵ -perturbation of a critical point. This locally-constant (at precision ϵ) set of local coordinates can be projected out, then, yielding an equivalent model with maximal-rank Jacobian in the neighborhood of this point (with lower-bounded associated minor determinant), with effective latent space $\tilde{L}$ of lower dimension than L .

We give two simple examples that imply our condition on certain limit points not existing in L :

1. U does not contain a boundary at all, e.g., a sufficiently strong regularity condition is imposed on $\mathcal{F}^{\mathbf{r}}$ such that it is a strong submersion on an unbounded subspace. This is similar to the first two examples previously mentioned.

2. $\mathcal{F}^r$ is a proper open map (onto its image) on $U \cup \partial U$, such as if it is composed of linear transformations and leaky rectified linear units. This is as under such a map, ∂U maps to $\partial(\mathcal{F}^r(U))$, which has no intersection with $\mathcal{F}^r(U)$ as $\mathcal{F}^r$ is open. To see this, assume some point $d \in \partial U$ mapped to $y \in \mathcal{F}^r(U)$. By the definition of this set, there must also be some $u \in U$ mapping to y under $\mathcal{F}^r(U)$. As $\mathcal{F}^r(U)$ is proper, there must be a neighborhood of y with preimage (intersected with U) compactly contained in U . However, the image of any Cauchy sequence (x_i) in U approaching $d \notin U$ must be in this neighborhood of y for sufficiently large i by continuity, violating compactness and yielding a contradiction.

Finally, we define a helper bump function that we use in our proof. Consider the smooth transition function:

$$g(x) := \left(\frac{\exp(-x^{-1})}{\exp(-x^{-1}) + \exp(-(1-x)^{-1})} \right). \quad (77)$$

The function

$$h_{a,b,c,d}(x) := g\left(\frac{x-a}{b-a}\right) g\left(\frac{d-x}{d-c}\right) \quad (78)$$

is smooth, equal to 1 on the interval $[b, c]$, and vanishes outside of the interval (a, d) . The periodic bump function (the “bubble wrap” function):

$$t(x) := \sum_{i=-\infty}^{\infty} h_{2\pi i, \frac{\pi}{2}+2\pi i, \pi+2\pi i, \frac{3\pi}{2}+2\pi i}(x) \quad (79)$$

is then a smooth function equal to 1 when $x \equiv z \pmod{2\pi}$ for all $\frac{\pi}{2} \leq z \leq \pi$, and equal to 0 when $x \equiv z \pmod{2\pi}$ for all $\frac{3\pi}{2} \leq z \leq 2\pi$.

With the preliminaries in place, we now prove our expressivity separation on the (ℓ, n, k) -HSMT task.

Theorem B.3 (Autoregressive hypergraph stabilizer measurement translation memory lower bound). *Consider an autoregressive model with C^2 contractible Finsler latent manifold L , and model function after $m = \binom{n}{k} + n$ tokens:²*

$$\mathcal{F} : \mathbb{R}^m \rightarrow L \times \mathbb{R}^n. \quad (80)$$

Consider as well $K \subset \mathbb{R}^m$, the space of $\mathbf{Q} = (\mathbf{v}, \boldsymbol{\gamma})$ where all $v_i = t(\|\boldsymbol{\gamma}\|_2^2)$, with t defined as in Eq. (79).

Assume that the projection of $\mathcal{F}(U)$ onto L is C^2 for some contractible, open subset $U \subseteq K$, where U is a C^1 manifold of dimension at least $\dim(L) + 2$. Assume also that $\mathcal{F}$ is a strong submersion on U , and that no sequence with limit point in $\partial U \not\subseteq U$ maps to a sequence with limit point in L under $\mathcal{F}$ projected onto L .

This model cannot achieve a finite backward empirical cross entropy on the $(\binom{n}{k} + 3n - k + 2, n, k)$ -hypergraph stabilizer measurement translation task due to quantum contextuality.

Proof. Due to the bijection between $\mathbf{Q} \in K$ and the associated hypergraph adjacency tensor $\mathbf{A}$ (i.e., $\boldsymbol{\gamma}$ is the vectorized $\mathbf{A}$), we will often use $\mathbf{A}$ to refer to its associated $\mathbf{Q}$ by proxy. It is obvious from this construction that K is an $\binom{n}{k}$ -dimensional embedding of representations of k -uniform hypergraphs via their adjacency tensors. Thus, the states described by points in K are exactly k -uniform weighted hypergraph states with, depending on the measurement results, perhaps overall phases on the stabilizers.

We first show that $\mathcal{F}$ induces a fiber bundle. We use $\Pi_L \circ \mathcal{F}$ to denote the projection of $\mathcal{F}$ onto L , and $\Pi_L \circ \mathcal{F}|_U$ its restriction to U . By Theorem 5.2 of Ref. [80], following our assumptions, $\Pi_L \circ \mathcal{F}|_U$ is

²Recall from Appendix B.1 that the first $\binom{n}{k} + n$ tokens are uniquely described by $(\mathbf{v}, \boldsymbol{\gamma}) \in \mathbb{R}^m$.

a fiber bundle with base space C^1 -diffeomorphic to L and induces a C^1 -diffeomorphism $\eta : H \times L \rightarrow U$ such that:

$$H \times L \cong U \quad (81)$$

for some fiber space H . In particular, $\Pi_L \circ \mathcal{F}|_U$ is a trivial bundle, and H is thus contractible and of dimension:

$$\dim(H) = \dim(U) - \dim(L) \geq 2. \quad (82)$$

We claim that $\eta(H \times \{\lambda\})$ is unbounded for all $\lambda \in L$. To see this, first note that as η is a homeomorphism, the boundary of $\eta(H \times \{\lambda\})$ for all $\lambda \in L$ must be a subset of ∂U . Assume now that it is nonempty, i.e., that there existed a Cauchy sequence with limit point in $\partial H \times \{\lambda\}$ that mapped to a Cauchy sequence with limit point in ∂U under η . Then, a sequence with limit point in ∂U would have base point $\lambda \in L$, violating our assumption that no sequence with limit point in the boundary of U has limit point in L under $\Pi_L \circ \mathcal{F}$. Thus, $\eta(H \times \{\lambda\})$ has no boundary. In particular, as $\eta(H \times \{\lambda\})$ is contractible, it is not compact as closed manifolds cannot be contractible. As $\eta(H \times \{\lambda\}) \subset K$ has no boundary it is therefore unbounded.

We now show that each fiber contains a sequence describing trivial operations in the (ℓ, n, k) -HSMT task. To see this, fix some $\lambda \in L$. As $\eta(H \times \{\lambda\})$ is unbounded and path-connected, there exist points with arbitrary 2-norm modulo 2π in $\eta(H \times \{\lambda\})$; namely, there exists a point $(\mathbf{h}_0, \lambda) \in H \times L$ such that (mapped under η):

$$\|\gamma\|_2^2 \equiv 0 \pmod{2\pi} \quad (83)$$

and thus all

$$v_i = t \left(\|\gamma\|_2^2 \right) = 0. \quad (84)$$

This corresponds to trivial measurements in the translation task, fixing the measurement outcomes to all be +1. Note that λ is the image of $\mathbf{h}_0$ under $\Pi_L \circ \mathcal{F}$. Assuming the model performs the translation task correctly, this fixes the measurement outcomes of the first n measurements to all be +1 for *all* points mapping to λ under $\Pi_L \circ \mathcal{F}$, as the next n outputs are just recapitulations of these first n measurement outcomes and they must be consistent for all inputs mapping to λ .

We now show the existence of other, nontrivial sequences sharing a fiber with $(\mathbf{h}_0, \lambda)$. Recall that H is unbounded and path-connected, and thus there exists some choice of $\bar{v} \in \mathbb{R}^{\binom{n}{k}}$ and $\gamma^* \in \mathbb{R}$ such that, for any $\gamma \in [\gamma^*, \infty)$,³ there exists some point $\mathbf{Q} \in K$ with $\gamma_{\bar{v}} = \gamma$ that maps to $\lambda \in L$ under $\Pi_L \circ \mathcal{F}$. Furthermore, as $\dim(H) \geq 2$, we can assume WLOG that the associated $\mathbf{v} \not\equiv 0 \pmod{2\pi}$ by varying the other coordinate of $\eta(H \times \{\lambda\})$. Thus, there exist $\mathbf{h}', \mathbf{h}'' \in U$ mapping to λ that describe k -uniform weighted hypergraph states with some respective hyperedges $e'_{\bar{v}}, e''_{\bar{v}}$ differing by 1.

By Lemma B.1 (in the qubit setting) or Lemma B.2 (in the qumode setting), we have that there exists a measurement sequence (describable by the translation task) of length $n - k + 2$ that antidistinguishes these three associated states with certainty. By then performing this antidistinguishing measurement sequence, the output for one of these $\mathbf{h}_0, \mathbf{h}', \mathbf{h}''$ must then be incorrect. Thus, the model must obtain an infinite backward empirical cross entropy on these three sequences when followed by the antidistinguishing measurement sequence. $\square$

B.4 Expressivity Separation for Encoder-Decoder Models

Though autoregressive sequence models are perhaps conceptually the simplest as they directly map input tokens to output tokens, in practice encoder-decoder models outperform them [42, 63]. We now show that no encoder-decoder model (with similar technical assumptions as the autoregressive separation) can perform the introduced tasks to finite backward empirical cross entropy. The proof will be similar to that of Theorem B.3; however, as encoder-decoder models can see the entire input sequence at once, we do not directly have the freedom to choose the antidistinguishing measurement sequence as in the proofs of those theorems. This will change the details of the proofs here.

³Or $\gamma \in (-\infty, -\gamma^*]$, though the logic proceeds identically.

We consider an encoder-decoder model with structure given by Fig. 1(b). The encoder of such a model is a function:

$$\mathcal{E}^{\mathbf{r}} : (\mathbb{R}^d)^\ell \rightarrow L \quad (85)$$

where, as in the proofs of Theorem B.3, $\mathbf{r}$ is a random vector such that, for any $\mathbf{r}$, $\mathcal{E}^{\mathbf{r}}$ is deterministic. For the reasons discussed in Appendix B.3 we will take the $\mathbf{r}$ -dependence to be implicit WLOG. We once again assume that L is a C^2 contractible Finsler manifold.

As $\mathcal{E}$ is a function of the *entire* input sequence, we have to appropriately modify our assumptions on its derivatives. We once again assume the existence of a contractible, $(\dim(L) + 2)$ -dimensional, open subset $U \subseteq K$, where K once again is a subspace of inputs corresponding to the measurement of hypergraph state stabilizers. To match the full sequence length, though, we now have to assume that for $\mathcal{E}$ on $U \times \{\mathbf{v}\}$ —not just U , as in the autoregressive model separation—for all but isolated $\mathbf{v} \in \mathbb{R}^{\ell - \dim(K)}$, $\mathcal{E}$ is a strong submersion, and no sequence with limit point in $\partial U \times \{\mathbf{v}\} \not\subset U \times \{\mathbf{v}\}$ maps to a sequence with limit point in L under $\mathcal{E}$.

With the preliminaries in place, we now prove our expressivity separation for encoder-decoder models.

Theorem B.4 (Encoder-decoder hypergraph stabilizer measurement translation memory lower bound). *Consider an encoder-decoder model with C^2 contractible Finsler latent manifold L , and encoder function*

$$\mathcal{E} : \mathbb{R}^{m+n+2^k+1} \rightarrow L, \quad (86)$$

where $m = \binom{n}{k} + n$. Let $K \subset \mathbb{R}^m$ be as in Theorem B.3. Consider any $V \subseteq \mathbb{R}^{n+2^k+1}$ that equals $\mathbb{R}^{n+2^k+1}$ up to the subtraction of isolated points. Assume there exists a contractible, open subset $U \subset K$, where U is a C^1 manifold of dimension at least $\dim(L) + 2$ such that for all $\mathbf{v} \in V$, on $U \times \{\mathbf{v}\}$, $\mathcal{E}$ is C^2 and a strong submersion. Finally, assume no sequence with limit point in $\partial U \times \{\mathbf{v}\} \not\subset U \times \{\mathbf{v}\}$ maps to a sequence with limit point in L under $\mathcal{E}$.

This model cannot achieve a finite backward empirical cross entropy on the $(\binom{n}{k} + 2n - k + 3, n, k)$ -hypergraph stabilizer measurement translation (without repetition) task due to quantum contextuality.

Proof. Consider $R := K \times \mathbb{R}^{n+2^k+1}$, with points of the form $(\mathbf{Q}, \mathbf{P})$ with $\mathbf{Q} \in K$ and $\mathbf{P} \in \mathbb{R}^{n+2^k+1}$. Just as in the proof of Theorem B.3, it is obvious from this construction that K is an $\binom{n}{k}$ -dimensional embedding of representations of k -uniform hypergraphs via their adjacency tensors. Thus, the states described by points in K are exactly k -uniform weighted hypergraph states with, depending on the measurement results, overall phases on the stabilizers.

We will proceed as follows. First, we will show that when $\dim(L)$ is sufficiently small, $\mathcal{E}$ must map three distinct $(\mathbf{Q}, \mathbf{P})$, $(\mathbf{Q}', \mathbf{P})$, and $(\mathbf{Q}'', \mathbf{P})$ to the same point in latent space $\lambda \in L$ for some fixed $\mathbf{P} = \mathbf{v}_0 \in V$. Then, we will use Lemma B.1 (in the qubit setting) or Lemma B.2 (in the qumode setting) to show that the stabilizers of these states give rise to an antidistinguishing measurement sequence of length $n - k + 2$ once the phases on their stabilizers have been fixed to be identical. Finally, we will show that one can find $(\mathbf{Q}, \mathbf{P}')$, $(\mathbf{Q}', \mathbf{P}')$, and $(\mathbf{Q}'', \mathbf{P}')$ also mapping to L , where $\mathbf{P}'$ is the associated antidistinguishing measurement sequence. This forces an incorrect output on one of these inputs, implying an infinite backward empirical cross entropy on any finite set containing these three measurement sequences.

We first show that $\mathcal{E}$ induces a fiber bundle. Let $\mathcal{E}|_{U \times \{\mathbf{v}\}}$ denote $\mathcal{E}$ restricted to $U \times \{\mathbf{v}\} \subset R$ for any choice of $\mathbf{v} \in V$. By Theorem 5.2 of Ref. [80], following our assumptions, $\mathcal{E}|_{U \times \{\mathbf{v}\}}$ is a fiber bundle for all $\mathbf{v} \in V$ with base space C^1 -diffeomorphic to L and induces a C^1 -diffeomorphism $\eta_{\mathbf{v}} : H_{\mathbf{v}} \times L \rightarrow U \times \{\mathbf{v}\}$ such that:

$$H_{\mathbf{v}} \times L \cong U \times \{\mathbf{v}\} \quad (87)$$

for some fiber space $H_{\mathbf{v}}$. In particular, $\mathcal{E}|_{U \times \{\mathbf{v}\}}$ is a trivial bundle, and $H_{\mathbf{v}}$ is thus path-connected and of dimension:

$$\dim(H_{\mathbf{v}}) = \dim(U) - \dim(L) \geq 2. \quad (88)$$

By the transitivity of C^1 -diffeomorphisms, all H_v are C^1 -diffeomorphic and can be taken to be some H WLOG. We thus have the trivial bundle structure:

$$H \times L \times V \cong U \times V, \quad (89)$$

with fibers C^1 -diffeomorphic to $H \times V$ and base space C^1 -diffeomorphic to L . We claim that $\eta_{v_0}(H \times \{\lambda\} \times \{v_0\})$ is unbounded for all $\lambda \in L, v_0 \in V$. To see this, first note that as η_{v_0} is a homeomorphism, the boundary of $\eta_{v_0}(H \times \{\lambda\} \times \{v_0\})$ for all $\lambda \in L$ must be a subset of $\partial U \times \{v_0\}$ for all $v_0 \in V$. Assume now that it is nonempty, i.e., that there existed a Cauchy sequence with limit point in $\partial H \times \{\lambda\} \times \{v_0\}$ that mapped to a Cauchy sequence with limit point in $\partial U \times \{v_0\}$ under η_{v_0} . Then, a sequence with limit point in $\partial U \times \{v_0\}$ would have base point $\lambda \in L$, violating our assumption that no sequence with limit point in the boundary of $U \times \{v_0\}$ has limit point in L under $\mathcal{E}$. Thus, $\eta_{v_0}(H \times \{\lambda\} \times \{v_0\})$ has no boundary. In particular, as $\eta_{v_0}(H \times \{\lambda\} \times \{v_0\})$ is contractible, it is not compact as closed manifolds cannot be contractible. As $\eta_{v_0}(H \times \{\lambda\} \times \{v_0\}) \subset K \times \{v_0\}$ has no boundary it is therefore unbounded.

We now show that each fiber contains a sequence beginning with m trivial operations in the (ℓ, n, k) -HSMT task. To see this, fix $\lambda \in L, v_0 \in V$. As $\eta_{v_0}(H \times \{\lambda\} \times \{v_0\})$ is unbounded and path-connected, there exist points with arbitrary 2-norm modulo 2π in $\eta_{v_0}(H \times \{\lambda\} \times \{v_0\})$; namely, there exists a point $(h_0, \lambda, v_0) \in H \times L \times \{v_0\}$ such that (mapped under η_{v_0}):

$$\|\gamma\|_2^2 \equiv 0 \pmod{2\pi} \quad (90)$$

and thus all

$$v_i = t \left(\|\gamma\|_2^2 \right) = 0. \quad (91)$$

This corresponds to trivial measurements in the translation task, fixing the measurement outcomes to all be $+1$. Note that λ is the image of (h_0, v_0) under $\mathcal{E}$. Assuming the model performs the translation task correctly, this fixes the measurement outcomes of the first n measurements to all be $+1$ for all points mapping to λ under $\mathcal{E}$ as the decoder $\mathcal{D}$ only depends on λ .

We now show the existence of other, nontrivial sequences sharing a fiber with (h_0, λ, v_0) . Recall that $H \times \{v_0\}$ is unbounded and path-connected, and thus there exists some choice of $\bar{v} \in \mathbb{R}^{\binom{n}{k}}$ and $\gamma^* \in \mathbb{R}$ such that, for any $\gamma \in [\gamma^*, \infty)$,⁴ there exists some point $(Q, v_0) \in K \times \{v_0\}$ with $\gamma_{\bar{v}} = \gamma$ that maps to $\lambda \in L$ under $\mathcal{E}$. Furthermore, as $\dim(H) \geq 2$, we can assume WLOG that the associated $v \not\equiv 0 \pmod{2\pi}$ by varying the other coordinate of $\eta_{v_0}(H \times \{\lambda\} \times \{v_0\})$. Thus, there exist $(h', v_0), (h'', v_0) \in U \times \{v_0\}$ mapping to λ that describe k -uniform weighted hypergraph states with some respective hyperedges $e'_{\bar{v}}, e''_{\bar{v}}$ differing by 1.

By Lemma B.1 (in the qubit setting) or Lemma B.2 (in the qumode setting), we have that there exists a measurement sequence (describable by the translation task) of length $n - k + 2$ that antidistinguishes these three associated states with certainty; by considering the dimensions of b, ϕ , and θ as described in Appendix B.1, this defines the first $n + 2^k$ coordinates of P . As V is $\mathbb{R}^{n+2^k+1}$ up to potentially the removal of isolated points, by varying the final coordinate—corresponding to measurements *after* the first $n - k + 2$ measurements—any given antidistinguishing measurement of length $n - k + 2$ can be assumed to be in V . As the construction given in Eq. (89) is unchanged when taking η_v rather than η_{v_0} as described, we can take v to be this antidistinguishing measurement sequence while still ensuring the associated three states map to the same point in the latent space of the model. Therefore, the output for one of these h_0, h', h'' must then be incorrect. Thus, the model must obtain an infinite backward empirical cross entropy on these three sequences when followed by the antidistinguishing measurement sequence. $\square$

⁴Or $\gamma \in (-\infty, -\gamma^*]$, though the logic proceeds identically.

C CV k -HRNNs as a Quantization of Recurrent Neural Networks

We assume the dimension of the output tokens y_i is $m = 1$ for notational simplicity in what follows; the generalization to larger m is immediate. The construction of CV k -HRNNs described in Sec. 3 is the quantization of alternating the classical Hamiltonian dynamics under the Hamiltonians:

$$G(\mathbf{x}_i) := \gamma(\mathbf{x}_i) \prod_{w_j \in \alpha(\mathbf{x}_i)} q_{w_j}, \quad (92)$$

$$H(\mathbf{x}_i) := \left(\sum_{j=1}^k \phi_j p_{v_j} + \sum_{\bar{w} \in 2^{\bar{v}}} \theta_{\bar{w}} \prod_{w_j \in \bar{w}} q_{w_j} \right) p_0, \quad (93)$$

where in the definition of H :

$$\beta(\mathbf{x}_i) = \bar{v}, \quad (94)$$

$$\kappa(\mathbf{x}_i) = (\phi, \theta) \quad (95)$$

and p_0 is the momentum of the ancillary output register used to perform phase estimation in Fig. 2.

Consider now discrete time dynamics under this Hamiltonian evolution with fresh ancillary modes labeled by 0 at each time step. We assume that initially all canonical variables are 0, other than an initial moment $p_0 = 1$ for the ancillary mode added at each time step. After evolution by G followed by H , we have from Hamilton's equations of motion the new canonical variables (given by primed notation):

$$p'_a = p_a - \delta_{a \in \alpha(\mathbf{x}_i)} \gamma(\mathbf{x}_i) \prod_{w_j \neq a \in \alpha(\mathbf{x}_i)} q_{w_j} - \sum_{\bar{w} \in 2^{\bar{v}}: a \in \bar{w}} \theta_{\bar{w}} \prod_{w_j \neq a \in \bar{w}} q_{w_j}, \quad (96)$$

$$q'_a = q_a + \phi_a, \quad (97)$$

$$q'_0 = \sum_{a=1}^k \phi_a \left(p_a - \delta_{a \in \alpha(\mathbf{x}_i)} \gamma(\mathbf{x}_i) \prod_{w_j \neq a \in \alpha(\mathbf{x}_i)} q_{w_j} \right) + \sum_{\bar{w} \in 2^{\bar{v}}} \theta_{\bar{w}} \prod_{w_j \in \bar{w}} q_{w_j}, \quad (98)$$

where δ denotes the indicator function.

In particular, a classical recurrent neural network [66] obeying these alternating Hamiltonian dynamics updates its latent vector as:

$$\lambda' = \mathbf{f}_{k-1}(\lambda; \mathbf{x}), \quad (99)$$

where $\mathbf{f}_{k-1}(\lambda; \mathbf{x})$ is some degree- $(k-1)$ polynomial in the entries of λ . Similarly, it outputs:

$$\mathbf{y} = \mathbf{g}_{k-1}(\lambda; \mathbf{x}), \quad (100)$$

where $\mathbf{g}_{k-1}$ is some degree- $(k-1)$ polynomial in the entries of λ .

D Time Complexity

We here discuss the implications of the existence of a polynomial-overhead classical simulation algorithm of the qumode k -HRNN as discussed in Sec. 3.1; we assume k is constant with respect to n in what follows.

As mentioned in the main text, our achieved memory separation is asymptotically optimal by Ref. [45], which gives an $O(n^k)$ -memory classical simulation algorithm for the quantum model we consider. This is accomplished via efficient representations of systems with dynamics constrained to belong to a low-dimensional Lie group; this is exactly our setting, where the dimension of the Lie group is $\Theta(n^k)$. As this algorithm requires multiplications of $\Theta(n^k)$ -dimensional representations of

members of this algebra, it achieves a time complexity of $O(n^{\omega k})$ per unit cell of the k -HRNN, where $\omega \geq 2$ is the matrix multiplication exponent (currently known to be at most 2.37187 [96], but for reasonable n no better than $\log_2(7) \approx 2.8$ using Strassen’s algorithm [97]). For $k = 2$ it is further known that $O(n^2)$ -time classical simulation algorithms exist for a unit cell of a k -HRNN [98, 99].

As discussed in Sec. 3.1, we expect training times for k -HRNNs to be at least $\Omega(n^k)$ per training iteration; this is required to even resolve the gradients of the model [33, 34, 62, 100]. Though the classical algorithm of Ref. [45] does not quite reach this scaling, this is only a constant-degree polynomial overhead, not an arbitrary-degree polynomial separation.

Interestingly, this gives a nice way to train the quantum model: one can classically train a k -HRNN via the classical simulation algorithm of Ref. [45] with only a cubic overhead, but still leverage an arbitrary-polynomial resource advantage by using a quantum computer at time of inference. In this way, one can consider the k -HRNN as a “quantum compression” of a quantum-inspired classical neural network model of size N trained on a classical computer. Once training is finished, one implements the model on a quantum computer using $\sim N^{\frac{1}{k}}$ qumodes. All future evaluations of the model now use polynomially fewer resources than the classical implementation, with the degree of separation depending on k . This also opens the possibility for the quantum model to “take over” training when in the vicinity of a minimum, where the loss landscape is believed to be more amenable for training on a quantum device [30, 31]. We leave further investigation of the implications of hybrid quantum-classical training algorithms for future work.

E Trainability in the Qubit-Based Construction

We here discuss a generalization of the qubit-based k -HRNNs considered in Sec. 3 which allows for efficient training of the model when k is constant. We here assume that the model is initialized with latent state $|\lambda_1\rangle = |+\rangle^{\otimes n}$.⁵ In each unit cell, a gate of the form:

$$C^{k-1} Z_{\bar{v}}(\phi) = \exp(-i\phi |1\rangle \langle 1|_{\bar{v}}) \quad (101)$$

is applied, and is followed by phase estimation of an operator of the form:

$$G_{\bar{v}}(\phi, \theta) := \exp\left(-i \sum_{i=1}^k \phi_i X_{v_i}\right) \exp\left(-i \sum_{\bar{w} \in 2^{\bar{v}}} \theta_{\bar{w}} |1\rangle \langle 1|_{\bar{w}}\right). \quad (102)$$

Here, we have used the shorthand:

$$|1\rangle \langle 1|_{\bar{v}} := \prod_{v_i \in \bar{v}} \left(\frac{I_{2^n} - Z_{v_i}}{2}\right). \quad (103)$$

We now present a modification of this construction which, though it is no longer translationally invariant, will allow for an efficient training algorithm while maintaining our proven classical lower bound for simulating the network.

We consider two predetermined time steps—that is, token indices— $\tau_2 \geq \tau_1$. We constrain our model such that:

1. For each time step indexed by $t \in [\tau_1]$, all $\kappa(\mathbf{x}_t)$ are constrained to be $\mathbf{0}$.
2. For each time step indexed by $t \in \{\tau_1 + 1, \dots, \tau_2\}$, all $\gamma(\mathbf{x}_t)$ are constrained to be 0 and all $G_{\bar{v}}(\kappa(\mathbf{x}_t))$ of the form:

$$G_j(\kappa(\mathbf{x}_t)) = \exp(-i\pi |1\rangle \langle 1|_j) \quad (104)$$

for j labeling a single qubit. We assume exactly $n - k$ distinct j are measured for time steps $t \in \{\tau_1 + 1, \dots, \tau_2\}$.

⁵Other initial latent states can be considered by fixing some number of initial tokens inputted to the k -HRNN.

3. For each time step indexed by $t > \tau_2$, all $\gamma(\mathbf{x}_t)$ are constrained to be 0.

Though this construction seems arbitrary, the qubit k -HRNN is constructed in such a way such that:

1. Time steps through $t \in [\tau_1]$ prepare a qubit hypergraph state.
2. Time steps $t \in \{\tau_1 + 1, \dots, \tau_2\}$ measure $n - k$ qubits in the computational basis.
3. Time steps $t > \tau_2$ measures operators of the form of G on the remaining qubits.

As quantum hypergraph states have a classically efficient stabilizer representation (see Appendix A.1), for k constant, all steps of this procedure are efficient to classically simulate. Just as is the case in the qumode setting, then—see Sec. 3.1—there are no quantum trainability barriers (QTBs) as the model can be simulated classically. Finally, our memory lower bound for classical simulation of the model still holds, as the subset of inputs $\mathbf{x}_i = (\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \gamma_i, \boldsymbol{\kappa}_i)$ we consider in the (ℓ, n, k) -HSMT task in our proofs of Theorems B.3 and B.4 obey these same constraints.