Quantum Doubly Stochastic Transformers

Jannis Born
Filip Skogh **Kahn Rhrissorrakrai**
Filippo Utro **Nico Wagner** **Aleksandros Sobczyk***

IBM Research

Correspondence to: jab@zurich.ibm.com

Abstract

At the core of the Transformer, the softmax normalizes the attention matrix to be right stochastic. Previous research has shown that this often de-stabilizes training and that enforcing the attention matrix to be doubly stochastic (through Sinkhorn's algorithm) consistently improves performance across different tasks, domains and Transformer flavors. However, Sinkhorn's algorithm is iterative, approximative, non-parametric and thus inflexible w.r.t. the obtained doubly stochastic matrix (DSM). Recently, it has been proven that DSMs can be obtained with a parametric quantum circuit, yielding a novel quantum inductive bias for DSMs with no known classical analogue. Motivated by this, we demonstrate the feasibility of a hybrid classical-quantum doubly stochastic Transformer (QDSFormer) that replaces the softmax in the self-attention layer with a variational quantum circuit. We study the expressive power of the circuit and find that it yields more diverse DSMs that better preserve information than classical operators. Across multiple small-scale object recognition tasks, we find that our QDSFormer consistently surpasses both a standard ViT and other doubly stochastic Transformers. Beyond the Sinkformer, this comparison includes a novel quantum-inspired doubly stochastic Transformer (based on QR decomposition) that can be of independent interest. Our QDSFormer also shows improved training stability and lower performance variation suggesting that it may mitigate the notoriously unstable training of ViTs on small-scale data.

1 Introduction

The Transformer [1] continues to be a dominant building block in natural language processing [2], computer vision [3, 4] and biology [5]. Quantum computing (QC), instead, is a novel paradigm with the potential to become practically useful in ML [6–10] and fuel applications across disciplines [11, 12]. Many attempts have been made to build Transformers with quantum gates, either entirely [13–15] or only the attention blocks [16–18]. However, rather than merely migrating, recent work in quantum ML identified constraints in specific flavors of neural networks (NN) and successfully mitigated those through quantum – e.g., fourier NNs [19], graph NNs [20] or input-convex NNs [21]. Some known limitations of Transformers are due to the softmax in the attention block, e.g., entropy collapse [22], rank collapse [23], token uniformity [24], eureka moments [25] and more [26–30]. Applying softmax enforces the attention matrices to be right-stochastic (i.e., rows sum to 1) while its temperature controls the distribution entropy and is often adjusted to stabilize training [25, 23].

*The author contributed to this work while at IBM Research.

Spotlight at 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Concurrently, it was discovered that Transformer attention naturally converge to doubly stochastic matrices (DSMs) over training, i.e. their rows *and* columns sum to 1 [31]. Motivated by this, the *Sinkformer* [31] enforces bistochasticity which boosts Transformer performance across different modalities (text, images, point clouds). Intuitively, doubly stochastic attention has a similar effect to increasing temperature (entropy) – attention becomes more "democratic", less interactions are missed and all tokens are being attended more equally. The Sinkformer [31] is a generalization of Transformers that leverages Sinkhorn's algorithm (SA) and has been widely adopted and extended [32–34]. Among various techniques to obtain DSMs [35–38], SA is the most obvious choice, however it has some disadvantages:

1. It is an iterative approximation procedure which reaches a DSM only in the limit. It is thus empirical how many iterations a Sinkformer needs and poor initialization can drastically deteriorate performance [39].
2. It can guarantee to find a DSMs only if the input matrix is non-negative, which is generally not the case within a Transformer (in practice non-negativity is enforced via exponentiation but we show that this hampers expressivity).
3. Backpropagating through SA often yields ill-conditioned and exploding/vanishing gradients when ε is small. In practice, under early stopping, SA is a sublinearly convergent mirror-descent fixed-point solver rather than a simple, well-conditioned layer [40].
4. It is non-parametric. Thus, in contrast to e.g., a NN layer, it cannot be optimized regarding *which* DSM should be returned.

Given the empirical superiority of the Sinkformer to vanilla Transformers, it is natural to study different techniques to make attention doubly stochastic. Strikingly, it was recently proven (in a

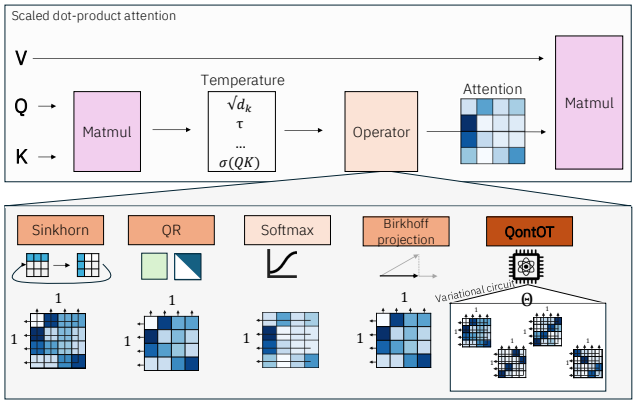


Figure 1: Doubly Stochastic Transformers. Standard scaled dot-product attention applies a Softmax activation on the query-key matrix (*top*). We study different techniques to make attention doubly stochastic attention by replacing the softmax operation (*bottom*). Our proposed Quantum Doubly Stochastic Transformer (QDSFormer) leverages QontOT, a variational quantum circuit with high expressivity.

different context) that DSMs can be obtained naturally with a variational (i.e., parametric) quantum circuit, dubbed QontOT [21]. They emphasize that there exists no classical *learning* (i.e., parametric) method that can produce a DSM, akin to QontOT. Here, we demonstrate that this opens the door for a hybrid quantum-classical doubly-stochastic Transformer (QDSFormer) which offers more flexibility than the Sinkformer. To that end, we extend QontOT to emit DSMs for an equally-sized matrix. The resulting quantum layer may replace the softmax within any standard (i.e., non-local, non-sparse) self-attention block. We focus on replacing the softmax inside the scaled dot-product self-attention of a Vision Transformer (ViT) for three reasons:

1. ViTs [3] suffer from unstable training [22, 25]
2. Unlike in NLP, the attention matrix size is constant, which eases quantum circuit application
3. The attention matrix in a Transformer encoder is unconstrained (unlike in decoders)

We empirically analyse expressivity of the quantum circuit, finding that it yields more diverse DSMs than Sinkhorn's algorithm both on synthetic and real data. It also preserves information better and induces higher entropy. We then train various flavors of doubly stochastic Transformers (see Figure 1) on more than ten object recognition datasets. In comparison to the ViT [3] and Sinkformer [31], the QDSFormer shows competitive performance, consistently surpassing both. In a compositional image

recognition task [25], we find that they stabilize Transformer training and accelerate learning as they antedate the Eureka moment in compositional problem solving.

In concurrent work, Shahbazi et al. have proposed the EPSFormer [34] and the LOTFormer [41], two doubly stochastic Transformers that, just like our QDSFormer, overcomes the dependence on Sinkhorn's algorithm to reach doubly stochastic attention. The ESPFormer [34] achieves this with sliced OT which is faster than SA but still slower than standard attention. Their improvement, the LOTFormer [41] marries doubly stochastic and linear attention via conditional OT, yielding better performance and scaling than softmax attention. Due to their concurrent nature, a performance comparison to ESPFormer and LOTFormer is not included in this work.

2 Methods

2.1 Doubly Stochastic Matrices (DSMs)

We denote the n-dimensional vector of ones by $\mathbf{1}_n$ and the $n \times n$ identity matrix as $\mathbf{I}_n$. The *Birkhoff polytope* $\Omega_n := \mathcal{N}(\mathbf{1}_n, \mathbf{1}_n)$ [42] defines the convex set of $n \times n$ doubly stochastic matrices (DSMs). A DSM $\mathbf{P} \in \Omega_n$ is a non-negative matrix with row/column sum of 1, i.e.,

$$\mathbf{P}\mathbf{1}_n = \mathbf{1}_n, \quad \mathbf{P}^\top \mathbf{1}_n = \mathbf{1}_n, \quad \mathbf{P}_{i,j} \geq 0. \tag{1}$$

A *right stochastic matrix* $\mathbf{R}$ has *row* sums of 1, i.e., $\mathbf{R}\mathbf{1}_n = \mathbf{1}_n, \quad \mathbf{R}_{i,j} \geq 0$, and a *left stochastic matrix* $\mathbf{L}$ has *column* sums of 1, i.e., $\mathbf{L}^\top \mathbf{1}_n = \mathbf{1}_n, \quad \mathbf{L}_{i,j} \geq 0$. Hence, a DSM is left and right stochastic. Moreover, the *Birkhoff-von Neumann theorem* states that the $n!$ vertices (i.e., extreme points) of the Birkhoff polytope Ω_n are permutation matrices, so their entries belong to $\{0, 1\}$. Notably, every DSM $\mathbf{P} \in \Omega_n$ can be decomposed as a convex combination of permutation matrices: $\mathbf{P} = \sum_{i=1}^{N} \lambda_i \mathbf{\Pi}_i$. Here $\boldsymbol{\lambda} \in \Delta_N$ is some probability vector in the probability simplex (denoted as Δ_N), $\{\mathbf{\Pi}_i\}$ are the $n \times n$ permutation matrices and $N \leq n^2$ denotes the extreme points. While the decomposition is not unique, each DSM can be represented by at most n^2 permutation matrices [43]. Due to the linear equality constraints, the Birkhoff polytope Ω_n lies within a $(n-1)^2$-dimensional affine subspace of the space of $\mathbb{R}^{n \times n}$ matrices.

2.2 Attention

We study extensions of dot product attention [1]

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{A}\mathbf{V} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\tau}\right)\mathbf{V} \tag{2}$$

where $\mathbf{Q} := \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} := \mathbf{X}\mathbf{W}_K$ and $\mathbf{V} := \mathbf{X}\mathbf{W}_V$ map the input $\mathbf{X}$ to query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$ through their respective weight matrix $\mathbf{W}_i$ s.t. $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{T \times d}$. Moreover, τ is called the "temperature" and canonically set to $\sqrt{d_k}$ [1]. It controls the entropy of the output: low temperature yields a peaky distribution emphasizing differences. High temperature attenuates differences thus increasing entropy. Note that $\tau^{-1}\mathbf{Q}\mathbf{K}^\top \in \mathbb{R}^{T \times T}$, so the unnormalized attention matrix is quadratic. Applying the softmax operator, denoted $S(\mathbf{z})_i = \frac{\exp(\mathbf{z}_i)}{\sum_{j=1}^{n} \exp(\mathbf{z}_j)}$, over the rows makes $\mathbf{A}$ right-stochastic, i.e., each row i contains a probability distribution denoting the amount of "attention" token i pays to the other tokens. The temperature τ,

2.3 Doubly-Stochastic Operators

Below we define a non-exhaustive set of operators that can transform $\mathbf{M} \in {\mathbb{R}_+}^{T \times T}$ to a DSM $\mathbf{P} \in \Omega_T$. The operators can be integrated into a Transformer by $\mathbf{M} := \mathbf{Q}\mathbf{K}^\top$ thus yielding a Doubly Stochastic Transformer ("DSFormer").

2.3.1 Sinkhorn's algorithm

The most natural approach to obtain a doubly stochastic Transformer was pursued in the Sinkformer [31] and leverages Sinkhorn's algorithm [35]. Sinkhorn's algorithm is based on Sinkhorn's

theorem, stating that for any square strictly positive matrix $\mathbf{M} \in {\mathbb{R}_+}^{T\times T}$, there exist (strictly) positive diagonal matrices $\mathbf{P} = \mathbf{D}_1, \mathbf{D}_2$ s.t., $\mathbf{D}_1\mathbf{M}\mathbf{D}_2 \in \Omega_T$. Sinkhorn's algorithm, also known as iterative proportional fitting [44], is an approximation procedure that iteratively normalizes the mass of the rows and the columns of $\mathbf{M}$ which has been proven to converge to a DSM by minimizing Kullback-Leibler (KL) divergence [45]. The sole hyperparameter of this procedure is K, the number of iterations, which we enforce to be odd, following [31], to ensure the resulting matrix is at least numerically row-stochastic, like for the canonical Softmax operator. Moreover, we study and compare two implementations of Sinkhorn's algorithm (SA), `Naive` and `OT`. `Naive` alternates between column- and row-normalization: at even iterations (t), each column is normalized as $\mathbf{P}_{ij}^{(t+1)} = \mathbf{P}_{ij}^{(t)} / \sum_i \mathbf{P}_{ij}^{(t)}$, and at odd iterations, each row is normalized as $\mathbf{P}_{ij}^{(t+1)} = \mathbf{P}_{ij}^{(t)} / \sum_j \mathbf{P}_{ij}^{(t)}$. Instead, the `OT` flavor is the operator used in the Sinkformer [31] which relies on the more robust and generalized version to compute optimal transport distances [37]. Note that, both flavors may not converge with few iterations, especially if $\mathbf{QK}^\top$ contains large numeric values. Therefore, the Sinkformer is only an approximately doubly stochastic Transformer.

2.3.2 Projection on the Birkhoff polytope

Previous work studied different approaches to project matrices onto the Birkhoff polytope [46, 47], and the most established scheme leverages Frobenius distance [48]. Alternatively, one can project $\mathbf{M}$ directly on Ω_T via $\mathbf{P} = \underset{\mathbf{X}\in\Omega_T}{\arg\min} \|\mathbf{X} - \mathbf{M}\|_F^2$, where the set for $\mathbf{X}$ and the objective are convex. We chose to minimize the Frobenius norm here but note that different distances could be explored. The resulting problem is a positive-definite convex quadratic program and can be rewritten as

$$\min_{\substack{s.t.\\ \mathbf{x}_i\geq 0\\ \mathbf{A}\mathbf{x}=\mathbf{1}_{2n}}} \frac{1}{2}\mathbf{x}^\top\mathbf{x} - \mathbf{q}^\top\mathbf{x}, \quad \mathbf{A} = \begin{pmatrix} \mathbf{1}_n^\top & \mathbf{0}^\top & \dots & \mathbf{0}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^\top & \mathbf{0}^\top & \dots & \mathbf{1}_n^\top \\ \mathbf{I}_n & \mathbf{I}_n & \dots & \mathbf{I}_n \end{pmatrix} \tag{3}$$

where $\mathbf{x} = \mathrm{vec}(\mathbf{X}^T)$, $\mathbf{q} = \mathrm{vec}(\mathbf{M}^T)$ and $\mathbf{A} \in \mathbb{R}^{2n\times n^2}$. The last (or any other) row of $\mathbf{A}$ can be removed without losing information, since it is a linear combination of the other rows ($\mathbf{A}$ has rank $2n-1$) [42, Thm. 8.1.1]. We solved the quadratic program with `OSQP` [49].

2.3.3 QontOT

QontOT is a parameterized (variational) quantum circuit that was conceived for conditional prediction of optimal transport plans [21] but can be extended to many combinatorial problems [50]. The circuit naturally emits DSMs and while [21] do not find signs of quantum advantage for their main task of optimal transport plan prediction, they report accuracy surpassing their classical baselines for the prediction of DSMs. This is likely a consequence of the choice of the ansatz which explores a previously unreported link between unitary operators and DSMs. Indeed, they first proved that DSMs can be obtained naturally with quantum computers thus constructing a quantum inductive bias for DSMs. Notably, as the authors state, it is currently unknown whether a similarly natural classical approach exists to produce DSMs *parametrically*.

Let $\odot$ be the Hadamard product and $\overline{U} = \left(U^\dagger\right)^\top$ the complex conjugate. For any unitary matrix $\mathbf{U}$: $\mathbf{U}\odot\overline{\mathbf{U}} \in \Omega_n$. Given the circuit parameters θ (typically in the hundreds) and $p \in \mathbb{R}$, QontOT obtains a DSM via $\mathbf{U}(p;\theta)\odot\overline{\mathbf{U}}(p;\theta)$. This matrix is block-decomposed before the classical rescaling. A notable detriment of QontOT is the data injection which is limited to a scalar p. Therefore, we extend the multiplicative data injection $f(\theta, p) = p\cdot\theta$ from scalars to tensors, such that $f(\theta, \mathbf{M}) = \theta \odot \overrightarrow{\mathbf{M}}$. If $\mathbf{M}$ has less items than θ we repeat its values to obtain a vector of length identical to θ. Furthermore, QontOT requires the DSM dimension n to be a power of 2. While this may be prohibitive within a Transformer (because sequence length T may differ), it can be mitigated by padding. Padding to powers of two is a common technique to maximize hardware efficiency. Here, we focus our experiments on ViTs because T is a function of patch size. In general, the circuit size scales favorably in $\mathcal{O}(\log_2(T))$. It needs at least $4(\log_2(T)+1)$ qubits, i.e., $2(q_d + q_a + 1)$ where q_d is the number of data qubits ($\log_2(T)$) and q_a is the number of auxilliary qubits ($\geq \log_2(T)+1$).

2.3.4 QR Decomposition

As highlighted above, any unitary $\mathbf{U}$ can provide a DSM by taking $\mathbf{U} \odot \overline{\mathbf{U}}$. For any input matrix $\mathbf{M}$, we can obtain a unitary $\mathbf{U}$ by computing an orthonormal basis for its column space. While there are many ways to obtain a basis, we choose a QR decomposition $\mathbf{M} = \mathbf{UR}$, in which case $\mathbf{R}$ is upper triangular. When implemented with Gram-Schmidt, QR is differentiable if $\mathbf{M}$ is full-rank, but for long-context applications that is rarely the case because query and key matrix have $d = \frac{d_{embed}}{n_{heads}}$ rows and typically $\mathbf{M} \in \mathbb{R}^{T \times T}$ has rank $\min\{d, T\}$, implying that $\mathbf{M}$ only has full rank when $d \geq T$. In practice, if the rank is defective, we inject additive Gaussian noise $\mathcal{N}(0, 1\mathrm{e}{-7})$ to obtain full ranks. In the ViTs studied in our experiments, $\mathbf{M}$ often has close to, or full-rank since the dimension d is greater than the number of patches P, where $P \approx T$. Moreover, QR has time complexity $\mathcal{O}(n^3)$ for dense $n \times n$ matrices, thus to scale up, approximation techniques may be needed [51].

3 Expressivity of Doubly-Stochastic Operators

Given the empirical superiority of the Sinkformer to the vanilla (i.e., right-stochastic) Transformer, a natural question is which operator to chose to obtain DSMs. Before training the DSFormers, we compare the expressivity of the operators – especially QontOT and Sinkhorn's algorithm – *in isolation* on synthetic data. We focus on two aspects.

1. **Soundness** – does the operator always produce a DSM? Given that $\mathbf{U} \odot \overline{\mathbf{U}} \in \Omega_n$, QontOT always yields a DSM. Similarly for the QR decomposition. Instead, Sinkhorn's algorithm (SA) may fail to produce a DSM if the input matrix is not positive. Within the Transformer where $\mathbf{M} \coloneqq \mathbf{QK}^\top$, the positivity requirement is generally not fulfilled which is mitigated by input exponentiation. Thus, following Sinkhorn's theorem, SA always converges *given* enough iterations k. But in practice the iterative procedure is limiting. When passing 8×8 $\mathbf{QK}^\top$ matrices from a trained Sinkformer, we observe that SA does *not* converge for the common choices of k (3 and 21). Indeed, the Frobenius distance to the Birkhoff polytope is $0.84_{\pm 0.3}$ for $k = 3$ and $0.23_{\pm 0.2}$ for $k = 21$. This is in contrast to the QR, QontOT and the Birkhoff projection which all yield distances $< 2\mathrm{e}{-4}$, QontOT even $< 5\mathrm{e}{-6}$. Instead, the vanilla Softmax operator yields a right-stochastic matrix with distance $1.12_{\pm 0.3}$. Hence Sinkformer attention is only approximately doubly stochastic.
2. **Completeness** – can the operator produce all possible DSMs? Sinkhorn's algorithm reaches all DSMs of the form $\mathbf{P} = \mathbf{D}_1\mathbf{M}\mathbf{D}_2$. However, due to the entry-wise exponentiation of $\mathbf{M}$ in the Sinkformer, the input matrix never contains any zero, thus the boundaries of the polytope cannot be reached. Regarding QontOT, the resulting DSM is a convex combination of unistochastic matrices [21, Eq. 11b]. Unistochastic matrices are a non-convex proper subset, covering a large amount of the Birkhoff polytope (albeit the exact amount is unknown [52]). In theory, if all unistochastic matrices could be reached, then by their convex combinations QontOT could cover the entire Birkhoff polyope. In practice, reaching all unistochastic matrices (especially all permutation matrices) with the same circuit parametrization is unfeasible as it requires fault-tolerant quantum hardware and high circuit depth (entanglement). But over the parameter space of QontOT, the Birkhoff polytope can be approximated more closely.

3.1 Empirical analysis

To empirically assess the completeness of the operators w.r.t. the Birkhoff polytope, we performed a brute-force analysis over a discretized grid of the unit hypercube. For a $n \times n$ matrix and a discretiztion step $d \in \mathbb{N}_+$, we sample each column from a discretized n-dimensional hybercube with d^n points, yielding d^{n^2} unique matrices. We refrain from analysing vectors above unit length because all operators are scale-invariant, i.e., $f(\lambda A) = f(A)$. For $n = 4$ and $d = 3$ we obtain $3^{16} \approx 43$M matrices and computed the DSM for each input, before rounding to third decimal place. Across all operators, QontOT yielded by far the most unique DSMs (see Figure 2**A**), behaving nearly injective when 8 or more circuit layers are used. This is important, because, none of the other operators is injective thus some information is lost when using it inside a neural network. A closer inspection of the empirical cumulative distribution function of all DSMs reveals that QR often emits the same DSMs and that with only 2 layers (i.e., 98 parameters), QontOT surpasses all other methods (see Appendix Figure A3). Furthermore, whereas all the classical techniques are non-parametric, QontOT yields a different set of DSMs for each parameter configuration. We repeated above experiment by

sampling from a discretized grid of the unit hypersphere (instead of the hypercube). In this case, Sinkhorn also produces collisions, while QontOT remains injective (there is no proof that the map is injective, but it appears to be experimentaly). With a discretization of $d = 3$, our sphere contains 625 matrices where all columns have unit lengths. Sinkhorn yields collisions by mapping all rank-1 matrices with constant rows to the center of the Birkhoff polytope, i.e., it fails to differentiate matrix $\mathbf{e}_2\mathbf{1}^\top$ and $\mathbf{e}_4\mathbf{1}^\top$ ($\mathbf{e}_2$ and $\mathbf{e}_4$ are the second and forth column of the identity) and thus produces only 621 unique DSMs whereas QontOT yields 625 DSMs (QR: 381). This is critical because it implies that Sinkhorn confuses cases where attention matrices are row-wise constant but each row has a unique value. In general, Sinkhorn's algorithm and the direct Birkhoff projection are both permutation and rotation equivariant. Instead, QR and QontOT do not possess this characteristic. Moreover, SA and the QR are scale-invariant whereas, again, QontOT is not. Note that such invariances or equivariances within a Transformer are not generally beneficial or detrimental.

Beyond approximate injectivity, a powerful operator needs to possess two further characteristics: First, information has to be preserved. Obtaining unique DSMs is useless if they destroy information from the input matrix. To assess this, we measured the Frobenius norm of the residuals between input

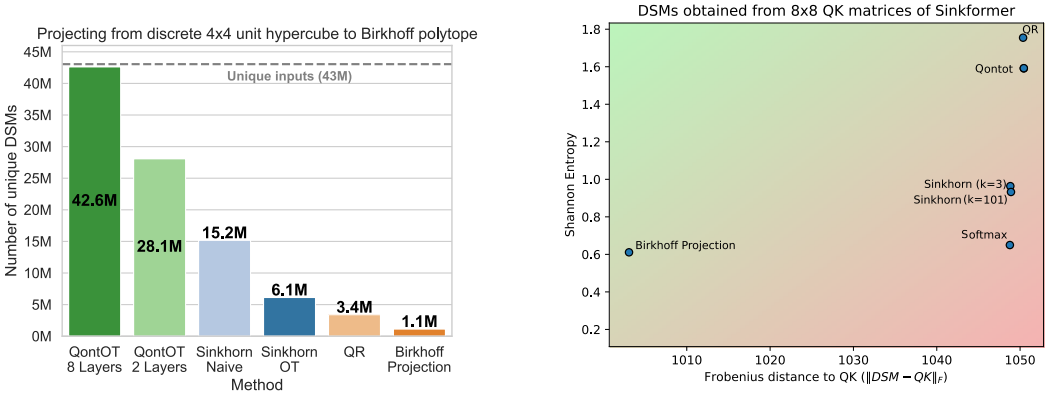


Figure 2: Left: Number of unique DSMs obtained after exhaustively iterating over a discretized unit hypercube. With only 8 layers, QontOT produces a unique DSM for every possible input, unlike all other methods. **Right**: Entropy vs. distance-preservation tradeoff. Shannon entropy of different doubly stochastic attention against the Frobenius norm of the difference between unnormalized attention $\mathbf{QK}$ and the obtained DSM $\mathbf{P}$.

and output matrix[2]. Secondly, low entropy has to be avoided because it causes vanishing gradients and destabilizes Transformer training [22, 28] for which various mitigation techniques have been suggested [53, 25]. This so-called "entropy collapse" arises if attention is too spiky and is induced by low temperature in the softmax. Our analysis in Figure 2**B** reveals that QontOT possesses comparable information preservation to Sinkhorn while having higher entropy on realistic unnormalized attention matrices. QR decomposition showed superior entropy, but its cubic scaling limits applicability beyond small-scale Transformers.

Next, we assessed which combination of circuit layers and auxiliary qubits yields the best speed-expressivity compromise. In general, a single circuit execution is in the three-digit millisecond range but can be efficiently parallelized. Increasing the number of layers causes a sub-linear runtime increase, whereas adding more qubits causes a exponential increase (Appendix Figure A2). Regarding expressivity, adding more layers has a higher impact than adding more qubits (Appendix Figure A4). In detail, we passed the same matrix $10,000$ times, sampled the circuit parameters $\theta \sim \mathcal{U}(-1, 1)$ and then measured the average range of values covered within each cell of the DSM. This shows that adding more auxiliary qubits is only useful if even more layers are added simultaneously.

4 Theoretical result on number of DSMs

The optimal way of studying the expressivity of a doubly-stochastic operator empirically would enumerate all DSMs in a given Birkhoff polytope Ω_n and assess for each DSM whether it can be reached (or how closely). The exact volume of the Birkhoff polytope is an open problem in

[2]Other metrics like measuring preservation of ranks or pairwise ratios are possible but yielded similar results.

mathematics [54] which limits our ability to study expressivity theoretically. In practice, one can assume a certain discretization $p \in \mathbb{N}_+$ s.t., $\mathbf{P}_{ij} \in \{0, \frac{1}{p-1}, ..., 1\}$, e.g., if $p = 2$ then $\mathbf{P}_{ij} \in \{0, 1\}$. In that case there are $n!$ DSMs. In Appendix E we provide a partial derivation for the combinatorial problem of identifying the function $f(n, p) \to \mathbb{N}$ returning the number of DSMs. The basic idea is that a $n \times n$ DSM has $(n-1)^2$ degrees of freedom, thus there are $p^{(n-1)^2}$ candidate matrices. Not all of these can be turned into DSMs because of two constaints, (1) the sum of any row or column must not exceed 1 and (2) the sum of the $n-1 \times n-1$ submatrix must not be below $n-2$ [54]. This allows to decompose f into $f(n, p) = p^{(n-1)^2} - c_1 - c_2 + c_{12}$ where c_1 and c_2 measure the constraint violations and c_{12} discounts cases where both constraints are violated. For details see Appendix E.

5 Quantum Doubly Stochastic Transformer

5.1 Experimental Setup

We evaluate different flavors of DSFormers obtained through replacing the Softmax function with any of the DSM operators described above. When integrating our QontOT-derived operator into a ViT we obtain our hybrid quantum-classical doubly stochastic Transformer (QDSFormer). In the following, we refer to QontOT as the attention flavor which contains the quantum circuit whereas QDSFormer denotes, more broadly, any Transformer with quantum doubly-stochastic attention. To date, the only realization of a QDSFormer is through QontOT. Among all operators, the classical ones are non-parametric whereas this quantum operator can be optimized during training. Therefore, one could theoretically optimize circuit parameters concurrently with Transformer training. However, the ViTs we study contains up to 4 attention layers, with a batch size of 512, yielding 2048 samples to optimize in a single forward pass. We predict, unless mentioned otherwise, 8×8 DSMs, use 16 circuit layers and 4 auxiliary qubits (16 qubits in total). Running the circuit on quantum hardware requires $\Omega(n^2/\varepsilon^2)$ shots to obtain satisfactory sampling error [21]. Assuming a precision of $\varepsilon = 0.01$, this is in the order of 640k shots per sample. Since quantum hardware operates on kHz frequency, execution and online optimization on hardware is unfortunately not (yet) feasible. Therefore, we perform exact statevector simulation with Qiskit [55] and implement three circuit training strategies:

1. **Differentiable**: Joint optimization of circuit and Transformer parameters through backpropagation, akin to integrating the circuit as a neural network layer. This is by far the slowest given the difficulty of gradient propagation through quantum circuits [56].
2. **Mixed**: A mixed strategy where Transformer training is interleaved with 200 steps of gradient-free circuit optimization with Nevergrad [57] on a per-epoch basis.
3. **Static**: The circuit is used in pure inference mode with parameters obtained from a 24-qubit DSM prediction experiment on quantum hardware [21].

From the operators studied in Section 3.1, we discard the Birkhoff projection due to its non-differentiability and low DSM-diversity (Figure 2**A**). For comparison, we further include the Norm-Softmax [58], here denoted as Softmax_σ, that attenuates attention by taking the minimum of the expected standard deviation $\tau := \sqrt{d_k}$ and the empirical one: $\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\min\left(\sigma(\mathbf{Q}\mathbf{K}^\top), \tau\right)}\right)$. This was found to stabilize ViT training [58, 25]. Moreover, replacing the standard deviation with the empirical variance, denoted as $\text{Softmax}_{\sigma^2}$, improved the performance and stabilized training even more. Note that both Softmax_σ and $\text{Softmax}_{\sigma^2}$ yields a right-stochastic but not a doubly-stochastic attention matrix. We did not perform hyperparameter optimization for any experiment (for details see Appendix D.1). We adapted the Sinkformer's ViT implementation of and simply reduced the number of layers and attention heads [31].

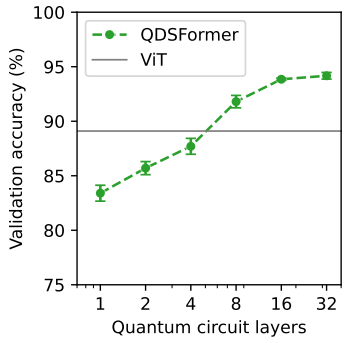
5.2 Data sets

We evaluate all ViTs on MNIST [59], Fashion MNIST [60], seven datasts from the MedMNIST benchmark [61] and a compositional task requiring multistep reasoning [25]. In that task, a 2×2 grid contains two MNIST digits (upper left and lower right) and two FashionMNIST items (upper right and lower left). If the digits have equal value, the label is the upper right fashion item, otherwise it is the bottom left fashion item. Performance typically ramps up quickly to $\sim 50\%$ because the model learns to attend one (and *only one*) of the FashionMNIST images. Upon continued training with a long saturation phase, a ViT suddenly grasps the relationship of the MNIST digits to the classification

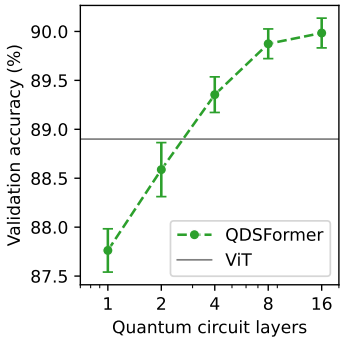
task and then climbs rapidly to a $90-95\%$ accuracy. The moment of abrupt improvement is called "Eureka moment" [25]. The dataset is split into 60K (10K) training (validation) examples. To accommodate the 8×8 attention matrix, each image from MNIST, FashionMNIST and MedMNIST is split into 7 horizontal stripes and a `CLS` token is pre-pended to the patch sequence. For the Eureka dataset we use a patch size of 14×28 pixels and mean-pooling.

5.3 Empirical results

First, we compare the QDSFormer directly with a standard ViT. The ViT uses softmaxed attention whereas the QDSFormer employs a ViT with a static (i.e., non-trainable) instantiation of QontOT to make attention doubly stochastic. Figure 3 clearly indicates that on both datasets, the QDSFormer exceeded the ViT by a significant margin. This confirms the finding that doubly stochastic attention



(a) 1-ViT-Layer on MNIST.



(b) 2-ViT-Layer on FashionMNIST.

Figure 3: Comparison of ViT and QDSFormer while varying the circuit depth. Mean/std from 5 trainings are shown. Within (**a**) and (**b**) all models use the same number of trainable parameters.

can improve ViTs [31]. Moreover, in both cases, adding more circuit layers increases performance logarithmically and with 4 or 8 circuit layers the ViT performance is surpassed. Exact numerical results are provided in Appendix Table A3. Next, we varied the number of ViT layers between 1 and 4, comparing to softmaxed attention, Softmax$_{\sigma^2}$ [25] and two classical doubly stochastic attention types: Sinkhorn as used in the Sinkformer [31] and a QR decomposition, a quantum-inspired alternative to QontOT. All flavors used the same number of parameters and training steps. On FashionMNIST (Table 1 *left*) the QDSFormer exceeded all other models for 2, 3 and 4 ViT layers with a performance delta larger than the standard deviation. The same result was obtained on MNIST and, this time, QontOT outperformed softmaxed attention also for one ViT layer (see Table 1 *right*).

Table 1: Validation accuracy of L-layered ViT on FashionMNIST and MNIST for different attention methods. QontOT uses 16 circuit layers. Mean/std computed from 5 trainings.

	FashionMNIST					MNIST				
L	**Softmax**	**Softmax$_{\sigma^2}$**	**QR**	**QontOT**	**Sinkhorn**	**Softmax**	**Softmax$_{\sigma^2}$**	**QR**	**QontOT**	**Sinkhorn**
1	$\underline{86.5}_{\pm 0.2}$	$75.3_{\pm 4.6}$	$\mathbf{87.1}_{\pm 0.3}$	$85.6_{\pm 0.1}$	$84.2_{\pm 3.6}$	$89.1_{\pm 12.5}$	$66.7_{\pm 22.5}$	$\mathbf{96.6}_{\pm 0.1}$	$93.9_{\pm 0.1}$	$\underline{94.3}_{\pm 2.0}$
2	$88.9_{\pm 0.1}$	$84.6_{\pm 2.1}$	$\underline{89.3}_{\pm 0.1}$	$\mathbf{90.0}_{\pm 0.2}$	$89.1_{\pm 0.7}$	$98.1_{\pm 0.3}$	$93.0_{\pm 4.6}$	$\underline{98.3}_{\pm 0.1}$	$\mathbf{98.4}_{\pm 0.1}$	$98.2_{\pm 0.3}$
3	$\underline{89.4}_{\pm 0.3}$	$86.3_{\pm 2.7}$	$\underline{89.4}_{\pm 0.1}$	$\mathbf{90.3}_{\pm 0.1}$	$\underline{89.4}_{\pm 0.8}$	$\underline{98.6}_{\pm 0.1}$	$97.7_{\pm 0.7}$	$\underline{98.6}_{\pm 0.1}$	$\mathbf{98.7}_{\pm 0.1}$	$\underline{98.6}_{\pm 0.1}$
4	$\underline{89.7}_{\pm 0.3}$	$87.1_{\pm 1.2}$	$89.5_{\pm 0.1}$	$\mathbf{90.3}_{\pm 0.1}$	$89.1_{\pm 1.1}$	$\mathbf{98.8}_{\pm 0.1}$	$97.9_{\pm 0.7}$	$\underline{98.7}_{\pm 0.1}$	$\mathbf{98.8}_{\pm 0.1}$	$97.9_{\pm 1.6}$

In further experiments with more ViT layers performance assimilated and plateaued due to the simplicity of the datasets. But we saw scant further improvement for more than 16 circuit layers. For a barplot visualization of Table 1 see Appendix Figure A5/A6. Notably, QontOT offers great flexibility in the type of ansatz for the quantum circuit [21]. We observed only minor differences between four different ansatz types, with three of them outperforming the ViT, underlining the generality of the finding (Appendix Table A1). A compelling aspect is that the static version of the QontOT-attention did perform as good or even better than the optimized one (see Appendix Figure A8). We tested an end-to-end optimizable QDSFormer where circuit and ViT parameters are jointly optimized. Such end-to-end training is not only slower, but also had lower accuracy than the static configuration, for both MNIST and FashionMNIST and 1 and 8 circuit layers (Figure A8). This may be caused

by Barren plateaus [62] (i.e., gradients are largely constant along most directions), a widespread phenomenon in variational quantum circuits that slows down learning. We further experimented with a "mixed" training strategy where the circuit is trained every n-th epoch. This did not reveal a clear benefit for more frequent circuit optimization (see Appendix Figure A7), potentially due to higher volatility of the circuit. We therefore use the static, faster configuration in all remaining experiments. Next, we repeated the comparison to the four classical attention types on larger datasets

Table 2: Test accuracy for MedMNIST datasets across 5 attention types in a 2-layer ViT.

MedMNIST dataset	Softmax	Softmax$_{\sigma^2}$	QR	QontOT	Sinkhorn
OCT	**64.4**$_{\pm 1.6}$	43.6$_{\pm 3.0}$	62.5$_{\pm 0.9}$	61.6$_{\pm 0.6}$	55.1$_{\pm 5.2}$
Pneumonia	84.2$_{\pm 0.8}$	84.7$_{\pm 2.0}$	84.3$_{\pm 0.7}$	**86.1**$_{\pm 1.0}$	83.0$_{\pm 1.5}$
Tissue	60.0$_{\pm 0.2}$	49.4$_{\pm 1.2}$	59.0$_{\pm 0.1}$	**60.6**$_{\pm 0.1}$	56.9$_{\pm 2.0}$
OrganA	78.8$_{\pm 0.5}$	73.6$_{\pm 1.7}$	78.4$_{\pm 0.6}$	**81.2**$_{\pm 0.3}$	77.0$_{\pm 2.5}$
OrganC	79.8$_{\pm 0.5}$	71.7$_{\pm 7.3}$	79.6$_{\pm 0.3}$	**82.7**$_{\pm 0.5}$	79.7$_{\pm 1.0}$
OrganS	64.4$_{\pm 0.6}$	59.3$_{\pm 0.9}$	62.6$_{\pm 0.8}$	**68.1**$_{\pm 0.6}$	63.5$_{\pm 0.9}$
Breast	79.6$_{\pm 2.0}$	78.2$_{\pm 2.2}$	**81.3**$_{\pm 2.9}$	80.0$_{\pm 1.1}$	80.1$_{\pm 0.8}$
Mean	73.0	65.8	72.5	**74.3**	70.8

(up to 240k images) from the MedMNIST benchmark [61]. In 5 out of 7 datasets, the QDSFormer obtained significantly better results than all other methods (Table 2), with a mean accuracy increase of 1.3% compared to a standard ViT. Notably, none of the other attention types can improve upon the standard ViT. Another important advantage of the QDSFormer is its stabilizing effect. Among repeated training runs, the performance variation (i.e., test accuracy variance) is consistently lower than for all classical methods (e.g., Table 1 or Table 2). Notably, with 1 ViT-layer and softmax-based attention some trainings on MNIST failed to converge.

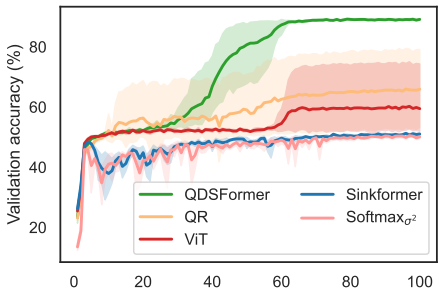
Furthermore, to study training stability more systematically, we used a compositional object recognition task with 10 classes, referred to as "Eureka" dataset. ViTs are very unstable to train on this task [25]. The random seed may determine whether the model saturated at 50% accuracy or experienced a Eureka moment (EM) after hundreds of epochs and would finally converge to 90% accuracy. As a mitigation strategy, Hoffmann et al. [25] tame the attention by replacing the Softmax with the NormSoftmax. In practice, temperature is often tuned manually to find a sweet spot between too low temperature (causing vanishing gradients by low entropy) and too high temperature (causing vanishing gradients by uniform attention). We speculated that doubly stochastic attention might, *en passant*, antedate the Eureka moment (EM) because it increases attention entropy without making it uniform [31], thus circumventing temperature tuning. Our experiments confirmed that the standard ViT implementation from [25] achieves its Eureka moment only after a few hundred epochs. While the same holds true for their proposed mitigation strategies (Softmax$_\sigma$ and Softmax$_{\sigma^2}$), the QDSFormer consistently learned within 100 epochs to solve this task, resulting in a 30% accuracy improvement over a standard ViT (Figure 4b). This major improvement, achieved with an extremely lightweight quantum circuit (1 layer) also consolidated across different learning rates (Figure 4a).

Next, to assess the potential of the QDSFormer beyond image data and simultaneously study the scalability of the QDSFormer, we applied it on a novel dataset for time-series classification of InfraRed (IR) spectra of molecules into 37 functional groups [63]. This dataset contains almost 1M samples and we scale up the circuit to produce 4x larger attention matrices (16×16). On this dataset, the performance differences are marginal and the QDSFormer performs on par with a standard ViT (for details see Appendix Table A2). This shows that the QDSFormer can meaningfully generalize to domains like scientific data and tasks like multi-label classification.

Since all above results were obtained through statevector simulation, we conducted a final experiment to understand the detrimental effect of quantum noise induced by real quantum hardware via the publicly available `IBM Quantum Platform`. We used the three machines *Torino* (Heron R1, 133 qubits, error per layered gate: 1.3%), *Brisbane* (Eagle R3, 127 qubits, EPLG: 2.2%) and *Cusco* (Eagle R3, 127 qubits, EPLG: 6.8%). This 14-qubit experiment tests the potential for a hybrid hardware training. Despite various light error mitigation techniques, the obtained doubly stochastic attention matrices consistently show high entropy (i.e., a tendency toward more uniform distributions), even for larger shot counts. Experimental details and plots are given in Appendix A). As Appendix Figure A1B

LR	Metric	Softmax	Softmax$_{\sigma^2}$	Sinkhorn	QR	QontOT
1e-3	Acc.	$53.4_{\pm 0.1}$	$48.6_{\pm 1.1}$	$49.0_{\pm 0.9}$	$\underline{61.2}_{\pm 11.0}$	$\mathbf{70.0}_{\pm 13.9}$
	EM@Ep	–	–	–	$\mathbf{72.2}_{\pm 14.2}$	$\underline{74.8}_{\pm 8.4}$
	# EM	0/5	0/5	0/5	**2/5**	**2/5**
7e-4	Acc.	$53.9_{\pm 0.2}$	$50.9_{\pm 0.8}$	$50.9_{\pm 0.7}$	$\underline{72.9}_{\pm 10.0}$	$\mathbf{82.3}_{\pm 14.0}$
	EM@Ep	–	–	–	$\underline{57.7}_{\pm 11.0}$	$\mathbf{43.3}_{\pm 7.0}$
	# EM	0/5	0/5	0/5	**4/5**	**4/5**
5e-4	Acc.	$61.1_{\pm 15.0}$	$51.0_{\pm 0.5}$	$51.6_{\pm 0.2}$	$\underline{66.4}_{\pm 16.0}$	$\mathbf{89.4}_{\pm 0.1}$
	EM@Ep	61.0	–	–	$\underline{72.6}_{\pm 22.0}$	$\mathbf{43.6}_{\pm 8.3}$
	# EM	1/5	0/5	0/5	$\underline{2/5}$	**5/5**

(a) Results on the Eureka dataset across different models and learning rates. EM@Ep denotes the average epoch of the EM; runs without EM are set to epoch 100.



(b) Validation accuracy per epoch, highlighting the Eureka Moment on the compositional dataset. Confidence bounds from 5 runs.

Figure 4: (a) Eureka results across attention methods. **(b)** QDSFormer antedates the Eureka Moment (EM).

shows, when comparing to the noise-free ground truth attention matrix, the ordering of the values in the attention matrix was preserved with high precision (spearman $\rho > 0.9$ even for moderate shot count). Since the circuit runs within a ViT, successfully preserving the ordering will be key (to not destroy signal). Instead, numerical exactness (cf. Appendix Figure A1A) may be compromised: embedding a noisy quantum attention block (which preserves peak attention scores but also increases entropy) into a Transformer could even be advantageous. The additional entropy may avoid vanishing gradients and act as a form of regularization. This effect is particularly notable compared to its noise-free analog, which remains classically intractable if sufficient qubits are used.

6 Conclusion

Here, we proposed the Quantum Doubly Stochastic Transformer. We conceived this method by connecting the centerpiece of a novel variational quantum circuit [21] with the Transformer, facilitated through the empirical observation that doubly-stochastic attention improves performance in Transformers [31]. By extending the QontOT circuit from scalars to matrices we enabled its integration into a ViT, thus providing the first parametric, doubly-stochastic Transformer. Notably, the QDSFormer presents a meta-class of Sinkformers because it estimates DSMs parametrically, i.e., it can be optimized to learn arbitrary transformations onto the Birkhoff polytope. Moreover, to our knowledge, there is no classical, parametric approach to estimate DSMs, thus the QDSFormer is a promising candidate for hybrid quantum-classical neural networks trained on quantum hardware.

Our empirical expressivity analysis revealed that the quantum circuit produces DSMs that are more diverse, preserve information better and have higher entropy than DSMs from Sinkhorn's algorithm. On multiple simple object recognition tasks, the QDSFormer exhibited significantly higher accuracy, outperforming a ViT and a Sinkformer in most cases. Our usage of quantum attention substantially stabilizes the notoriously unstable ViT training on small-scale data, as evidenced by the performance on a compositional object recognition task (Figure 4b), previously used to study ViT training dynamics [25]. Albeit these results are promising, all experiments were performed on comparably small-scale, due to the currently poor scaling of quantum computers in general (which is expected to improve). Notably, by leveraging QR decomposition, we also proposed a novel, quantum-*inspired* attention flavor. Broadly speaking, outsourcing the activation function to a parametric quantum circuit might be seen a computational overhead, however, we envision that this may reveal potential benefits (typically in small-data, small-model and short-training settings [10]) that are out of reach for classical hardware. To that end, future work could explore concurrent optimization of ViT and circuit parameters via the parameter-shift rule on real quantum hardware.

Funding Declaration

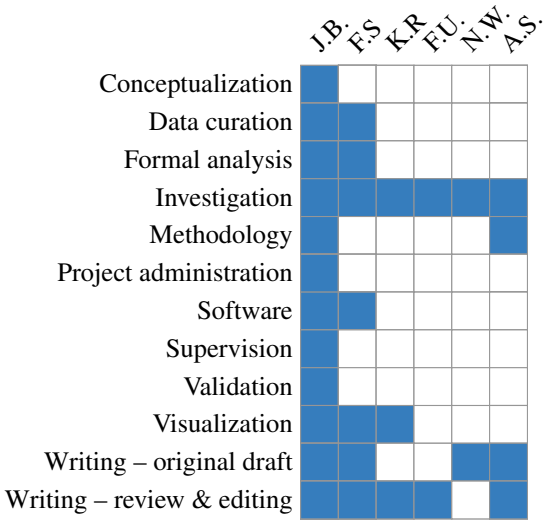
This project received no external funding.

Competing Interests

The authors declare no competing financial or non-financial interests.

Author Contributions.

Conceptualization: J.B.; **Data curation:** J.B., F.S; **Formal analysis:** J.B., F.S; **Investigation:** J.B., F.S, K.R, F.U., N.W., A.S.; **Methodology:** J.B., A.S.; **Project administration:** J.B.; **Software:** J.B., F.S; **Supervision:** J.B.; **Validation:** J.B.; **Visualization:** J.B., F.S, K.R; **Writing – original draft:** J.B., F.S, N.W., A.S.; **Writing – review & editing:** J.B., F.S, K.R, F.U., A.S.



References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, et al. Attention is all you need. *Advances in neural information processing systems*, 30(1):261–272, 2017. 1, 3

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. 1, 2

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[5] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024. 1

[6] Maria Schuld and Nathan Killoran. Is quantum advantage the right goal for quantum machine learning? *Prx Quantum*, 3(3):030101, 2022. 1

[7] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567 (7747):209–212, 2019.

[8] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 2021. ISSN 1745-2481.

[9] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. *Nature Communications*, 12(1):2631, 2021.

[10] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021. 1, 10

[11] Saugata Basu, Jannis Born, Aritra Bose, Sara Capponi, Dimitra Chalkia, Timothy A Chan, Hakan Doga, Mark Goldsmith, Tanvi Gujarati, Aldo Guzman-Saenz, et al. Towards quantum-enabled cell-centric therapeutics. *arXiv preprint arXiv:2307.05734*, 2023. 1

[12] Alberto Di Meglio, Karl Jansen, Ivano Tavernelli, Constantia Alexandrou, Srinivasan Arunachalam, Christian W Bauer, Kerstin Borras, Stefano Carrazza, Arianna Crippa, Vincent Croft, et al. Quantum computing for high-energy physics: State of the art and challenges. *PRX Quantum*, 5(3):037001, 2024. 1

[13] Nikhil Khatri, Gabriel Matos, Luuk Coopmans, and Stephen Clark. Quixer: A quantum transformer model. *arXiv preprint arXiv:2406.04305*, 2024. 1

[14] Xuan-Bac Nguyen, Hoang-Quan Nguyen, Samuel Yen-Chi Chen, Samee U Khan, Hugh Churchill, and Khoa Luu. Qclusformer: A quantum transformer-based framework for unsupervised visual clustering. *arXiv preprint arXiv:2405.19722*, 2024.

[15] Naixu Guo, Zhan Yu, Matthew Choi, Aman Agrawal, Kouhei Nakaji, Alán Aspuru-Guzik, and Patrick Rebentrost. Quantum linear algebra is all you need for transformer architectures. *arXiv preprint arXiv:2402.16714*, 2024. 1

[16] Iordanis Kerenidis, Natansh Mathur, Jonas Landman, Martin Strahm, Yun Yvonna Li, et al. Quantum vision transformers. *Quantum*, 8:1265, 2024. 1

[17] Ethan N Evans, Matthew Cook, Zachary P Bradshaw, and Margarite L LaBorde. Learning with sasquatch: a novel variational quantum transformer architecture with kernel-based self-attention. *arXiv preprint arXiv:2403.14753*, 2024.

[18] Eyup B Unlu, Marçal Comajoan Cara, Gopal Ramesh Dahale, Zhongtian Dong, Roy T Forestano, Sergei Gleyzer, Daniel Justice, Kyoungchul Kong, Tom Magorsch, Konstantin T Matchev, et al. Hybrid quantum vision transformers for event classification in high energy physics. *Axioms*, 13(3):187, 2024. 1

[19] Jiaming Zhao, Wenbo Qiao, Peng Zhang, and Hui Gao. Quantum implicit neural representations. In *Forty-first International Conference on Machine Learning*, 2024. 1

[20] Slimane Thabet, Mehdi Djellabi, Igor Olegovich Sokolov, Sachin Kasture, Louis-Paul Henry, and Loic Henriet. Quantum positional encodings for graph neural networks. In *Forty-first International Conference on Machine Learning*, 2024. 1

[21] Nicola Mariella, Albert Akhriev, Francesco Tacchino, Christa Zoufal, Juan Carlos Gonzalez-Espitia, Benedek Harsanyi, Eugene Koskin, Ivano Tavernelli, Stefan Woerner, Marianna Rapsomaniki, Sergiy Zhuk, and Jannis Born. Quantum theory and application of contextual optimal transport. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. 1, 2, 4, 5, 7, 8, 10, 19

[22] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023. 1, 2, 6

[23] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022. 1

[24] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2793–2803. PMLR, 18–24 Jul 2021. 1

[25] David T Hoffmann, Simon Schrodi, Jelena Bratulić, Nadine Behrmann, Volker Fischer, and Thomas Brox. Eureka-moments in transformers: Multi-step tasks reveal softmax induced optimization problems. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18409–18438. PMLR, 21–27 Jul 2024. 1, 2, 3, 6, 7, 8, 9, 10, 19, 25, 27

[26] Zhilin Yang, Thang Luong, Russ R Salakhutdinov, and Quoc V Le. Mixtape: Breaking the softmax bottleneck efficiently. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[27] Haw-Shiuan Chang and Andrew McCallum. Softmax bottleneck makes language models unable to represent multi-mode word distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1, 2022.

[28] Kai Shen, Junliang Guo, Xu Tan, Siliang Tang, Rui Wang, and Jiang Bian. A study on relu and softmax in transformer. *arXiv preprint arXiv:2302.06461*, 2023. 6

[29] Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, and Guanghui Wang. Accumulated trivial attention matters in vision transformers on small datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3984–3992, 2023.

[30] Shen Yuan and Hongteng Xu. Towards better multi-head attention via channel-wise sample permutation. *arXiv preprint arXiv:2410.10914*, 2024. 1

[31] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022. 2, 3, 4, 7, 8, 9, 10, 19, 27

[32] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[33] Jong Chul Ye, Yujin Oh, et al. Otseg: Multi-prompt sinkhorn attention for zero-shot semantic segmentation. In *The 18th European Conference on Computer Vision, ECCV 2024*. European Computer Vision Association (ECVA), 2024.

[34] Ashkan Shahbazi, Elaheh Akbari, Darian Salehi, Xinran Liu, Navid NaderiAlizadeh, and Soheil Kolouri. ESPFormer: Doubly-stochastic attention with expected sliced transport plans. In *Forty-second International Conference on Machine Learning*, 2025. 2, 3

[35] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964. 2, 3

[36] Nikitas Rontsis and Paul Goulart. Optimal approximation of doubly stochastic matrices. In *International Conference on Artificial Intelligence and Statistics*, pages 3589–3598. PMLR, 2020.

[37] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 4

[38] Fei Wang, Ping Li, and Arnd Christian Konig. Learning a bi-stochastic data similarity matrix. In *2010 IEEE International Conference on Data Mining*, pages 551–560. IEEE, 2010. 2

[39] James Thornton and Marco Cuturi. Rethinking initialization of the sinkhorn algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 8682–8698. PMLR, 2023. 2

[40] Flavien Léger. A gradient descent perspective on sinkhorn. *Applied Mathematics & Optimization*, 84(2): 1843–1855, 2021. 2

[41] Ashkan Shahbazi, Chayne Thrash, Yikun Bai, Keaton Hamm, Navid NaderiAlizadeh, and Soheil Kolouri. Lotformer: Doubly-stochastic linear attention via low-rank optimal transport. *arXiv preprint arXiv:2509.23436*, 2025. 3

[42] Richard A. Brualdi. *Combinatorial Matrix Classes*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2006. 3, 4

[43] Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A*, 5:147–154, 1946. 3

[44] Michael Bacharach. Estimating nonnegative matrices from marginal data. *International Economic Review*, 6(3):294–310, 1965. 4

[45] George W Soules. The rate of convergence of sinkhorn balancing. *Linear algebra and its applications*, 150:3–40, 1991. 4

[46] Derek Lim, René Vidal, and Benjamin D Haeffele. Doubly stochastic subspace clustering. *arXiv preprint arXiv:2011.14859*, 2020. 4

[47] Tianjiao Ding, Derek Lim, Rene Vidal, and Benjamin D Haeffele. Understanding doubly stochastic clustering. In *International Conference on Machine Learning*, pages 5153–5165. PMLR, 2022. 4

[48] Ron Zass and Amnon Shashua. Doubly stochastic normalization for spectral clustering. *Advances in neural information processing systems*, 19, 2006. 4

[49] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020. doi: 10.1007/s12532-020-00179-2. URL https://doi.org/10.1007/s12532-020-00179-2. 4

[50] Dylan Laplace Mermoud, Andrea Simonetto, and Sourour Elloumi. Variational quantum algorithms for permutation-based combinatorial problems: Optimal ansatz generation with applications to quadratic assignment problems and beyond. *arXiv preprint arXiv:2505.05981*, 2025. 4

[51] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011. 5

[52] Charles Dunkl and Karol Życzkowski. Volume of the set of unistochastic matrices of order 3 and the mean jarlskog invariant. *Journal of mathematical physics*, 50(12), 2009. 5

[53] Shulun Wang, Feng Liu, and Bin Liu. Escaping the gradient vanishing: Periodic alternatives of softmax in attention mechanism. *IEEE Access*, 9:168749–168759, 2021. 6

[54] Clara S Chan and David P Robbins. On the volume of the polytope of doubly stochastic matrices. *Experimental Mathematics*, 8(3):291–300, 1999. 7, 23

[55] Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D Nation, Lev S Bishop, Andrew W Cross, et al. Quantum computing with qiskit. *arXiv preprint arXiv:2405.08810*, 2024. 7

[56] Amira Abbas, Robbie King, Hsin-Yuan Huang, William J Huggins, Ramis Movassagh, Dar Gilboa, and Jarrod McClean. On quantum backpropagation, information reuse, and cheating measurement collapse. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[57] Pauline Bennet, Carola Doerr, Antoine Moreau, Jeremy Rapin, Fabien Teytaud, and Olivier Teytaud. Nevergrad: black-box optimization platform. *ACM SIGEVOlution*, 14(1):8–15, 2021. 7

[58] Zixuan Jiang, Jiaqi Gu, and David Z Pan. Normsoftmax: Normalizing the input of softmax to accelerate and stabilize training. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–6. IEEE, 2023. 7

[59] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 7, 25

[60] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747. 7, 25

[61] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195, 2021. 7, 9, 25

[62] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):4812, 2018. 9

[63] Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Advances in Neural Information Processing Systems*, 37:125780–125808, 2025. 9, 19, 25

[64] Nic Ezzell, Bibek Pokharel, Lina Tewala, Gregory Quiroz, and Daniel A Lidar. Dynamical decoupling for superconducting qubits: a performance survey. *Physical Review Applied*, 20(6):064027, 2023. 16

[65] Joel J Wallman and Joseph Emerson. Noise tailoring for scalable quantum computation via randomized compiling. *Physical Review A*, 94(5):052325, 2016. 16

[66] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106. 19

[67] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. 19

[68] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 19

[69] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 19

[70] Sumeet Khatri, Ryan LaRose, Alexander Poremba, Lukasz Cincio, Andrew T Sornborger, and Patrick J Coles. Quantum-assisted quantum compiling. *Quantum*, 3:140, 2019. 19

[71] Liam Madden and Andrea Simonetto. Best approximate quantum compiling problems. *ACM Transactions on Quantum Computing*, 3(2):1–29, 2022. 19

[72] Adam Smith, MS Kim, Frank Pollmann, and Johannes Knolle. Simulating quantum many-body dynamics on a current digital quantum computer. *npj Quantum Information*, 5(1):106, 2019. 19

A Quantum hardware experiment

We measured the extent and the effect of quantum hardware noise on a DSM produced by our quantum circuit. To that end, we picked a shallow circuit (1 layer) with 14 qubits and computed the ground-truth 8×8 DSM for a random input matrix through statevector simulation. We then transpiled the circuit on three different quantum computers (Cusco, Brisbane and Torino) available to the public via the `IBM Quantum Platform`. After using transpilation optimization level 1, we obtained a circuit with a 2-qubit-depth of 15 and and a total of 52 two-qubit gates. As error mitigation techniques, we used dynamical decoupling [64], Pauli twirling [65] and a projection to the Birkhoff polytope of the approximate-DSM obtained from the quantum circuit (see Section 2.3.2). The results, shown in Figure A1, indicate that, consistently, `Cusco` was the noisiest machine and `Torino` yielded the best results. Moreover, in general, beyond a shot count of $10,000$ little performance improvement can be observed. This is a positive finding because it is substantially below the theoretical minimum given by the shot noise limit $(640,000)$. However, the deviation from the exact DSM, measured in Frobenius Distance, was substantial (Figure A1**A**). We analyzed the root cause of this and found that the deviation can be largely attributed to an increase in entropy. DSMs obtained from noisy quantum hardware converge toward the center of the Birkhoff polytope ($1/n$ in every cell). The relative ordering of the absolute values instead is largely preserved (Figure A1**C**).

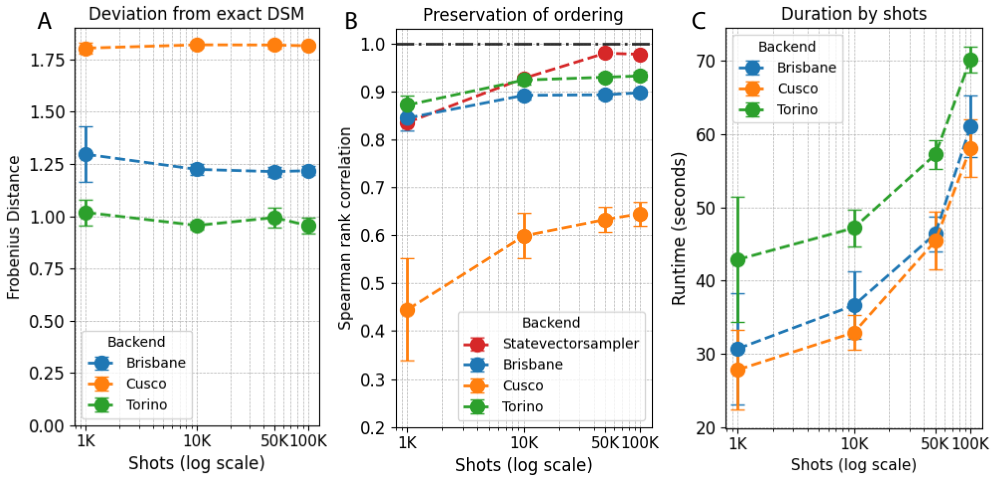


Figure A1: Hardware experiment on different quantum computers available via `IBM Quantum Platform`. **A** The Frobenius distance between the hardware-obtained DSM to its noise-free equivalent. **B**: The spearman rank correlation between the 64 values in the noise-free and hardware-obtained DSMs show that the ordering of values is largely preserved. Statevectorsampler here denotes finite sampling from an ideal, noise-free statevector.

B Circuit execution times

In Figure A2 we report detailed runtimes for the QontOT algorithm for different combinations of circuit layers and auxiliary qubits.

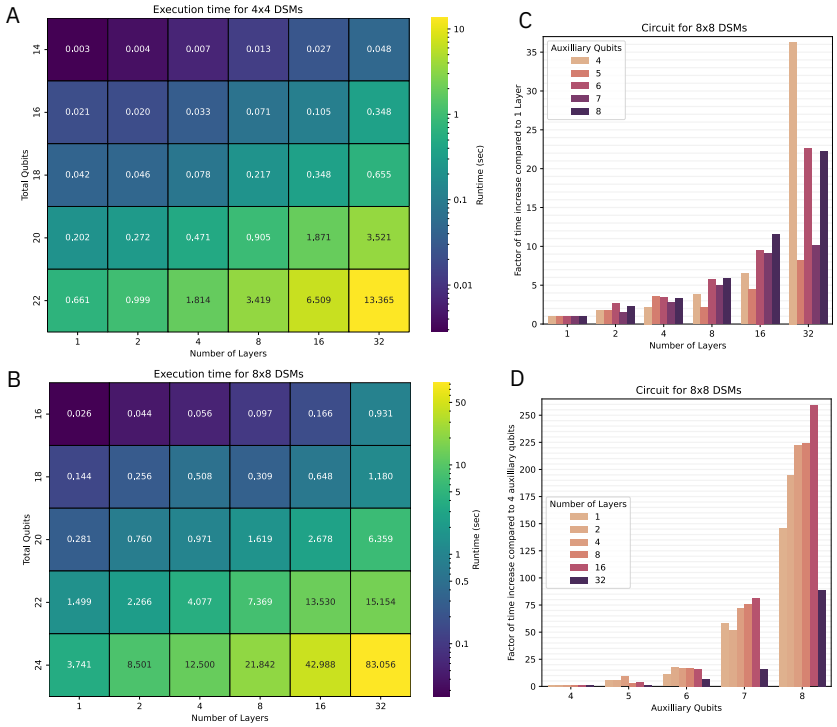


Figure A2: QontOT circuit execution times for DSM of size 4 (**A**) and 8 (**B**) for different combinations of qubits and circuit layers. **C** and **D** show the relative increase in execution time as a function of increasing the number of layers (**C**), and qubits (**D**). Adding more layers has a sublinear effect on runtime, adding qubits requires exponential more runtime. The minimal number of auxiliary qubits is $\log_2(n) + 1$ and the total number of qubits is $2(q_d + q_a)$ where q_d and q_a are data and auxilliary qubits respectively.

C Empirical circuit expressivity

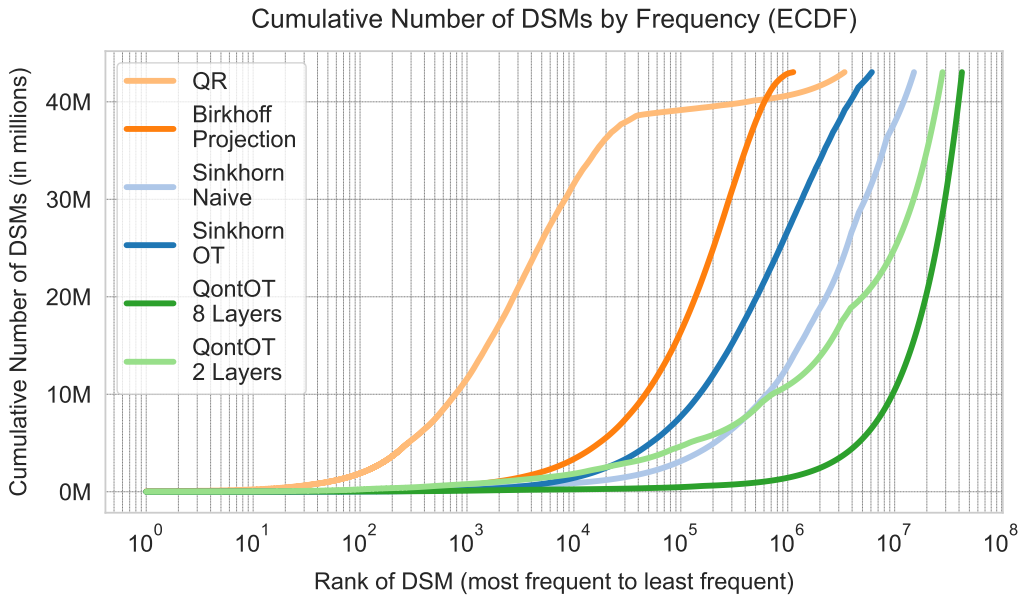


Figure A3: DSM counts were ranked descendingly and plotted against their cumulative count. QontOT generally produces more diverse DSMs than Sinkhorn's algorithm.

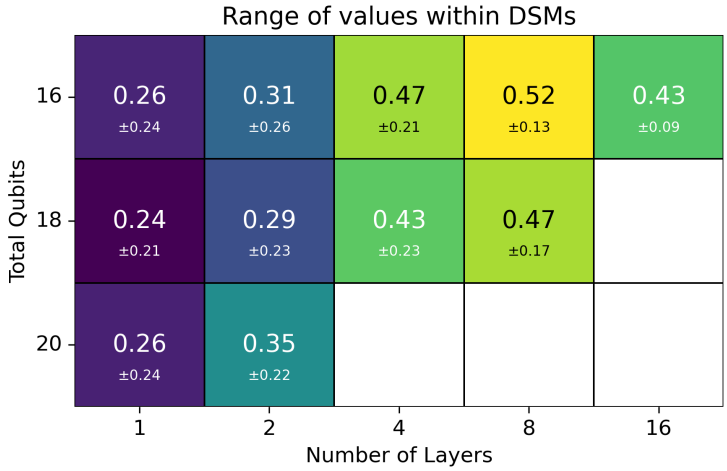


Figure A4: Mean range of observed values in the DSM obtained from a single, random input matrix, when randomly sampling 1000 circuit parametrizations.

D QDSFormer results

D.1 Hyperparameters

For experiments on MNIST, FashionMNIST and the seven MedMNIST datasets, the ViT was configured with a hidden dimension of 128 and an MLP dimension expansion factor of 1. The model was tested with 1 to 4 Transformer layers, each containing a single attention head. No dropout was applied, and the batch size was set to 100. For the optimizer, Adam [66] was used, and the learning rate schedule followed the setup in Sander et al. [31], with an initial learning rate of 5e-4, decreasing by a factor of 10 at epochs 31 and 45. For the more complex Eureka dataset, comprised of 56x56 RGB images, the hidden dimension was increased to 256, and a larger batch size of 512 was used. The MLP expansion factor was also doubled to 2. A cosine learning rate schedule was used with the optimizer AdamW [67]. The scheduler uses 5 warmup epochs with a warm-up learning rate of 1e-6, the decay rate is set to 0.1 and the minimum learning rate is 1e-5, the other parameters follows the default TIMM settings [68]. For the optimizer the weight decay is 0.05 and betas $(0.9, 0.999)$. All studied imaging datasets (MNIST, FashionMNIST and seven types of MedMNIST datasets) come with predefined train/validation/test splits. On the infrared spectral data of molecules from Alberts et al. [63] we performed a 5-fold cross validation with $80\%/20\%$ train/test split. Hyperparameters were kept identical to previous experiments. For the Eureka dataset, no Exponential Moving Average (EMA) is used. Experiments were conducted on `POWER8` infrastructure in `Python 3.9` with `PyTorch` [69] 1.13.1 on machines with 16 cores of `32GiB RDIMM DDR4 2.7 GHz`. Due to the small size of the ViTs, training took between few hours and a day (for the slowest, i.e., end-to-end-differentiable configuration of the QDSFormer). The Sinkformer [31] and the standard ViT implementation are taken from the original author's repository: `https://github.com/michaelsdr/sinkformers`. The results on the compositional Eureka dataset [25] were generated with the ViT implementation of the original authors: `https://github.com/boschresearch/eurekaMoments`. The implementation of the QontOT circuit was implemented as described in Mariella et al. [21] and, as described in the main text, adapted to digest matrix (or vector) inputs rather than scalars only.

D.2 QontOT ansatz types

Table A1: Ablation study for a 2-layer QDSFormer with different circuit ansatz types and varying number of layers on FashionMNIST. Mean/std of 5 runs.

Circuit L.	**Simple**	**Parted**	**Centrosymmetric**	**Trotter**
1	$\mathbf{88.0}_{\pm 0.10}$	$87.7_{\pm 0.22}$	$86.4_{\pm 0.23}$	$87.7_{\pm 0.08}$
8	$\mathbf{89.9}_{\pm 0.15}$	$89.8_{\pm 0.15}$	$89.4_{\pm 0.17}$	$88.4_{\pm 0.20}$

Simple: This ansatz is the most generic and resembles a checkerboard structure formed by 4-parameter unit-blocks acting on two qubits each [70, 71]. If all parameters are zero, it falls back to the identity. This ansatz is convenient because it is shallow in simulation but whose depth may vary depending on qubit layout of the quantum hardware.
Parted: This ansatz partitions the *Simple* ansatz into two parts: $U = U_1 \otimes U_2$, where U_2 operates normally, and U_1 is transposed and placed around the initial Bell state. This design reduces the original *Simple* ansatz circuit depth nearly by half, which may be more efficient on certain quantum hardware. However depending on the qubit layout of the quantum hardware, it carries the potential of increased transpiled circuit depth, as the two-qubit gates may act on distant qubits necessitating additional swap gates upon transpilation, which we observed on IBM Eagle and Heron quantum processing units. Unless mentioned otherwise, we used this ansatz in all our experiments as it yields shallower circuits in simulation and the increased depth compared to the *Simple* ansatz was negligible at tested system sizes.
Centrosymmetric: This was the predominantly used ansatz by Mariella et al. [21]. It is less generic, biasing toward properties of centrosymmetric matrices.
Trotter: This ansatz implements a second-order Trotter decomposition [72]. Each circuit layer corresponds to a Trotter step.

D.3 Time series classification

Table A2: Micro-F1 on IR spectra dataset across 5-fold cross-validation with a 1-layer ViT. QDSFormer uses 16 circuit layers for both DSM sizes.

DSM	Softmax	Softmax$_{\sigma^2}$	QR	QDSFormer	Sinkhorn
8×8	$81.60_{\pm 0.34}$	$81.41_{\pm 0.23}$	$\underline{81.68}_{\pm 0.18}$	$\mathbf{81.70}_{\pm 0.05}$	$81.38_{\pm 0.22}$
16×16	$\mathbf{81.55}_{\pm 0.07}$	$80.94_{\pm 0.13}$	$\underline{81.48}_{\pm 0.10}$	$81.06_{\pm 0.27}$	$80.98_{\pm 0.30}$

D.4 Ablation studies

Table A3: QDSFormer ablation varying the circuit layers. Exact numbers corresponding to Figure 3.

Configuration	Validation Accuracy (%) MNIST	FashionMNIST
QDSFormer-1L	$83.4_{\pm 0.73}$	$87.7_{\pm 0.22}$
QDSFormer-2L	$85.7_{\pm 0.60}$	$88.5_{\pm 0.27}$
QDSFormer-4L	$87.7_{\pm 0.73}$	$89.3_{\pm 0.18}$
QDSFormer-8L	$91.8_{\pm 0.57}$	$89.8_{\pm 0.15}$
QDSFormer-16L	$93.8_{\pm 0.10}$	$\mathbf{90.0}_{\pm 0.15}$
QDSFormer-32L	$\mathbf{94.2}_{\pm 0.30}$	$\mathbf{90.0}_{\pm 0.13}$
Baseline		
ViT	$92.9_{\pm 3.76}$	$88.9_{\pm 0.12}$

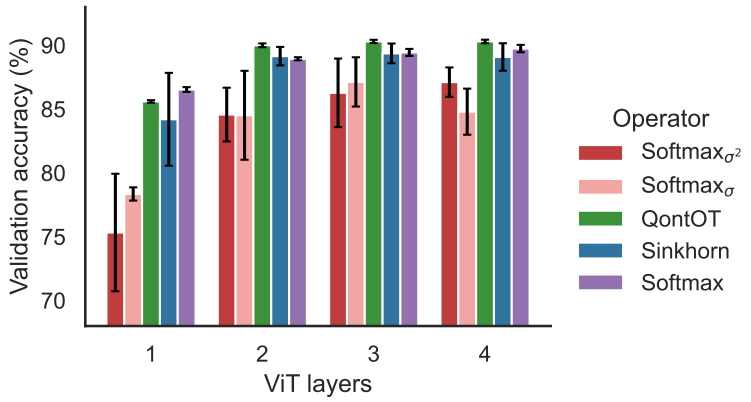


Figure A5: FashionMNIST results of different ViT layers for different attention types.

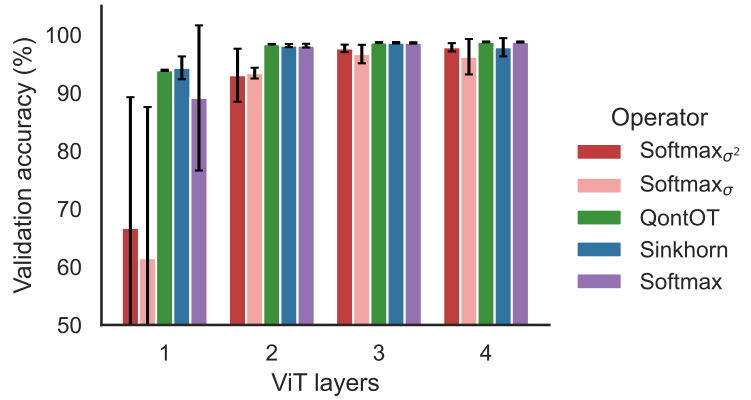


Figure A6: MNIST results of different ViT layers for different attention types.

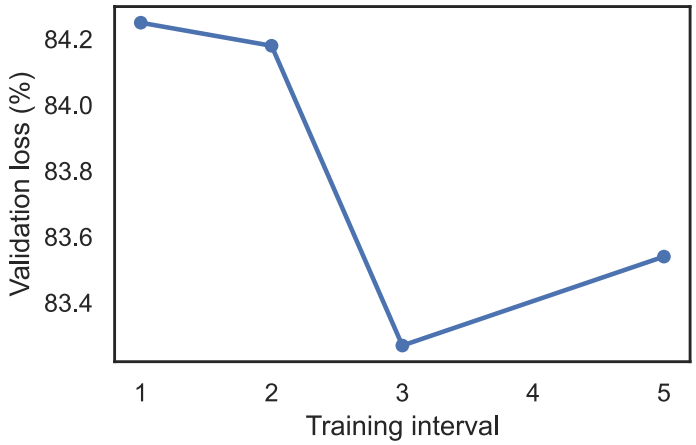
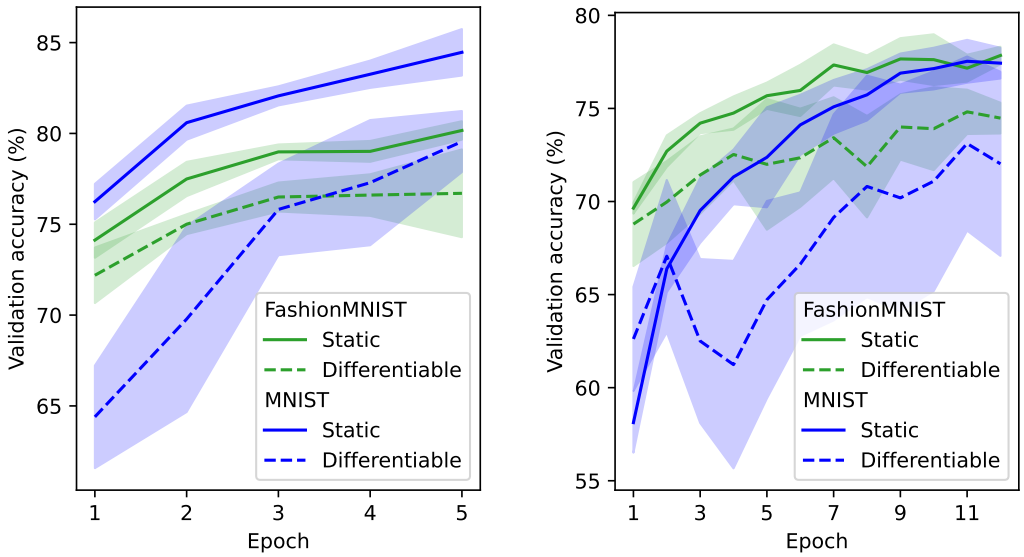


Figure A7: Ablation on different number of QontOT trainings on MNIST in the mixed circuit optimization strategy.

D.5 Differentiable circuit



(a) QDSFormer 8 circuit layers.

(b) QDSFormer 1 circuit layers.

Figure A8: Differentiable QontOT training versus the static circuit.

E Counting DSMs

In Section E.1 we provide a full analytical solution to count the number of 3×3 DSMs for a given discretization $p \in \mathbb{N}$. In Section E.2 we strive to extend the analytic solution to an arbitrary $n \in \mathbb{N}$ but only provide a partial solution. Finally, in Section D.3 we provide numerical results from brute-force calculation of the number of DSMs that verified the explicit analytical solution in Section E.1.

E.1 Analytical solution for $n = 3$

Intuition. By systematically testing all combinations from the discretized range of values, starting with an initial 2×2 zero matrix, each element is incrementally increased in the order $x_{0,0}, x_{1,0}, x_{0,1}, x_{1,1}$ with the next highest value in the discretized range. Once an element reaches its maximum value, the next element is increased, and the preceding elements are reset to 0. This cycle repeats, starting again with the first element.

Explanation. Assume that $n = 3$ and a specific discretization $p \in \mathbb{N}_+$ are given. In this scenario, the corresponding 3×3 matrix possesses 4 degrees of freedom, implying that the associated submatrix has dimensions 2×2. The first constraint requires that the sum of the elements in each row and each column of the matrix must not exceed 1.

If a specific element e_{ij} with $i, j \in \{0, 1\}$ is chosen and assigned a value x_i, the possible values for the remaining elements in the same row and column can be determined.

Given that each element can assume exactly p distinct values, the total number of combinations is computed as a sum over all p values:

$$f(3, p) = \sum_{i=1}^{p} c_i \tag{4}$$

The possible values for the elements in the same row and column are restricted to the subset $\{x_1, \ldots, x_{p-i+1}\}$. As a result, the amount of submatrices that satisfy the first constraint can be expressed as:

$$f(3, p) = \sum_{i=1}^{p} \left[\sum_{j=1}^{p-i+1} \left[\sum_{k=1}^{p-i+1} c_{ijk} \right] \right] \tag{5}$$

To determine the possible values for the last element, it is necessary to consider the elements e'_{ij}. The minimum number of possible values derived from these elements defines the number of candidates for the last element:

$$f(3, p) = \sum_{i=1}^{p} \sum_{j=1}^{p-i+1} \sum_{k=1}^{p-i+1} \sum_{l=1}^{\min(p-j+1, p-k+1)} \mathbb{1}(i, j, k, l, p) \tag{6}$$

Up to this point, only the first constraint has been considered. To fully satisfy the problem requirements, matrices that violate the second constraint must be excluded. The second constraint is satisfied when the sum of the indices of all elements does not exceed p. Instead of subtracting $1 - \mathbb{I}(i, j, k, l, p)$, the condition is captured using an indicator function $\mathbb{I}(i, j, k, l, p)$, defined as:

$$\mathbb{1}(i, j, k, l, p) = \begin{cases} 1 & \text{if } i + j + k + l - 3 \geq p, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

By incorporating $\mathbb{1}(i, j, k, l, p)$, the expression extends to:

$$f(3, p) = \sum_{i=1}^{p} \sum_{j=1}^{p-i+1} \sum_{k=1}^{p-i+1} \sum_{l=1}^{\min(p-j+1, p-k+1)} \mathbb{1}(i, j, k, l, p) \tag{8}$$

This can be summarized as follows:

$$f(3,p) = \sum_{(i,j,k,l)\in D(p)} \mathbb{1}(i,j,k,l,p) \tag{9}$$

where $D(p) = \{(i,j,k,l) \mid 1 \leq i \leq p, 1 \leq j \leq p-i+1, 1 \leq k \leq p-i+1, 1 \leq l \leq \min(p-j+1, p-k+1)\}$.

This equation has been validated computationally for values up to $p = 43$, demonstrating alignment with our empirical results.

E.2 General approximation

To determine the number of unique DSMs for a given $n, p \in \mathbb{N}$ we try to solve

$$f(n,p) = p^{(n-1)^2} - c_1 - c_2 + c_{12} \tag{10}$$

where the first term calculates the number of DSM-candidate matrices, c_1 and c_2 measure how often the constraints are violated and c_{12} is a small correction term counting cases where both constraints are violated.

Generally, c_{12} is very small, yet difficult to compute, thus a tight lower bound can be given with the remaining three terms. Below, we provide a derivation for c_2. We leave the derivation of c_1 to future work.

E.2.1 Constraint 2

Constraint 2. The sum of the $n-1 \times n-1$ inner matrix must not be below $n-2$ [54].

We aim to find a function $c_2(n,p)$ that computes the number of violations to Constraint 2 for a given $n, p \in \mathbb{N}$ when exhaustively looping over all $p^{(n-1)^2}$ candidate matrices that uniquely determine a $n \times n$ DSM.

An $n-1 \times n-1$ matrix where each cell x_{ij} can take p values has

$$|u| = (n-1)^2(p-1) + 1 \tag{11}$$

unique possible sums. These sums are regularly spaced from 0 to $(n-1)^2$ with a step size of $p-1$, i.e., $u_i = \left\{ \frac{i}{p-1} \mid i \in \{0, 1, ..., |u|\} \right\}$. This allows conversion to an integer problem (by multiplication of $p-1$) and apply *Stars & Bars Theorem 2*.

Theorem D.1. *For any $s, k \in \mathbb{N}$, the number of k-tuples $(x_0, ..., x_k)$ where $x_k \in \mathbb{N}_0$ with sum s is equal to the number of multisets of cardinality s taken from a set of size k:*

$$\binom{s+k-1}{k-1} \tag{12}$$

Specifically, we set $k := (n-1)^2$ and then define the set of sums that violate the constraint as $S := \{s \in \mathbb{N}_0 \mid 0 \leq s \leq (n-2)(p-1)\}$. Thus $|S| = (n-2)(p-1)$ We then compute the violations via:

$$\hat{c}_2(n,p) = \sum_{s=0}^{(n-2)(p-1)} \binom{s+(n-1)^2-1}{(n-1)^2-1} \tag{13}$$

Unfortunately, this is only approximately correct because Theorem D.1 assumes $x_k \in \mathbb{N}_0$, instead we require $x_i \in \{0, 1, \ldots, p-1\}$. Therefore, we exclude solutions where any $x_i > p-1$ through the inclusion exclusion principle as follows.

Assume that some $x_i > p-1$. We set $\hat{x}_i = x_i - (p-1+1)$. Since $\hat{x}_i \geq 0$, we can rewrite the original sum

$$s = \sum x_i = \sum_{i\in I} (\hat{x}_i + p) + \sum_{i\notin I} x_i \tag{14}$$

where I is the set of indices where $x_i > p - 1$.

Now let $m := |I|$ be the number of violating variables, then

$$s - mp = \sum_{i \in I} \hat{x_i} + \sum_{i \notin I} x_i \tag{15}$$

To find the number of non-negative integer solutions to Equation 15, we can again leverage Theorem D.1, but now in a corrected form:

$$d(n, p, m, s) = \begin{cases} \binom{s-mp+(n-1)^2-1}{(n-1)^2-1} & \text{if } s - mp > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

This gives a solution for a specific sum $s \in S$ and number of violations m. However, since $0 \leq m \leq (n-1)^2$, we have to sum over all options of m and apply the inclusion-exclusion principle to avoid over-/undercounting.

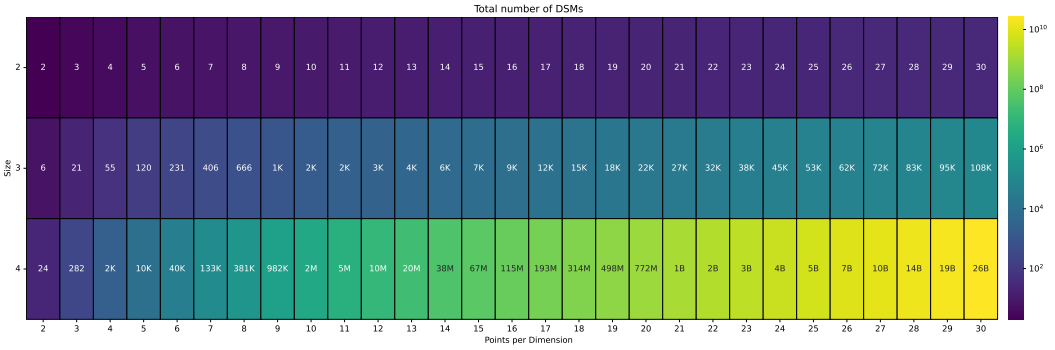


Figure A9: Number of DSMs of fixed size n with a given number of discretization steps p; up to values of $n = 4$ and $p = 30$.

$$c_2(n, p, s) = \sum_{m=0}^{(n-1)^2} \left[(-1)^m \binom{(n-1)^2}{m} d(n, p, m, s) \right] \tag{17}$$

where $\binom{(n-1)^2}{m}$ accounts for the number of ways to choose m out of $(n-1)^2$ variables that exceed $p - 1$.

Plugging Equation 17 back into the initial summation over all values $s \in S$ violating the constraint (see Equation 13), we obtain the final formula:

$$\begin{aligned} c2(n, p) &= \sum_{s=0}^{(n-2)(p-1)} c_2(n, p, s) \\ &= \sum_{s=0}^{|S|} \sum_{m=0}^{(n-1)^2} (-1)^m \binom{(n-1)^2}{m} \begin{cases} 0 & \text{if } s - mp \leq 0, \\ \binom{s-mp+n^2-2n}{n^2-2n} & \text{else.} \end{cases} \end{aligned}$$

where $|S| = (n-2)(p-1)$.

D.3 Empirical results

To empirically determine the solutions to $f(n, p)$ we implemented a brute-force algorithm by iterating over all $p^{(n-1)^2}$ candidate matrices of size $(n-1) \times (n-1)$ and verifying whether the two constraints are not violated (see section 4).

The results are given in Figure A10 and Figure A9. Interestingly, $f(n, 2) = n!$, but in general $f(n, p)$ scales super-factorially in n, for a given p.

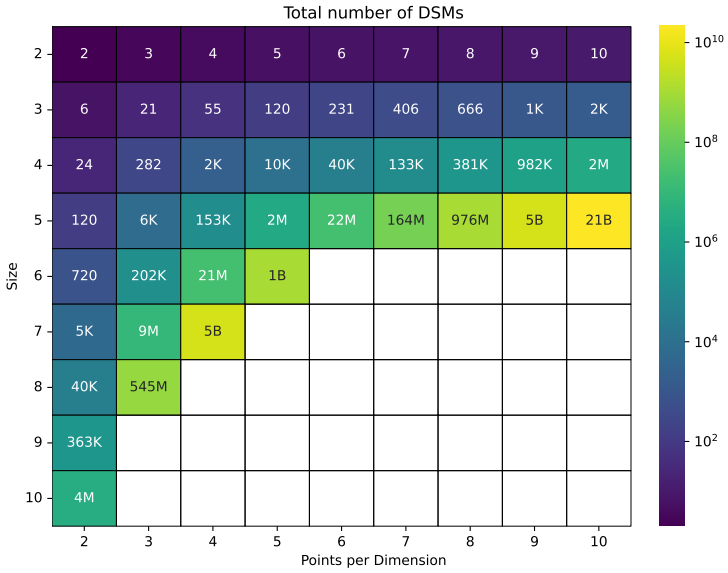


Figure A10: Number of DSMs of fixed size n with a given number of discretization steps p; up to values of $n = 10$ and $p = 10$. Empty cells require > 5 days of compute time on a machine with 128 cores and 128GB RAM.

E Checklist information

Dataset	Reference	License	Size
MNIST	[59]	GNU	70,000
Fashion-MNIST	[60]	MIT	70,000
OCTMNIST	[61]	CC BY 4.0	109,000
PneumoniaMNIST	[61]	CC BY 4.0	5,856
TissueMNIST	[61]	CC BY 4.0	236,386
OrganAMNIST	[61]	CC BY 4.0	58,830
OrganCMNIST	[61]	CC BY 4.0	23,583
OrganSMNIST	[61]	CC BY 4.0	25,211
BreastMNIST	[61]	CC BY 4.0	780
Compositional	[25]	GNU / MIT	70,000
IR Spectra	[63]	CDLA	790,000

Table A4: Summary of datasets used, with references, licenses, and sizes.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: We report empirical evidence for all claims in the abstract and introduction.
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The conclusion discusses limitations of our approach imposed by the noisy nature of current quantum hardware as well as the limited scaling to large-scale datasets.
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: In the appendix, we provide a full analytical solution to count the number of 3×3 DSMs for a given discretization. We then we strive to extend the analytic solution to an arbitrary $n \in \mathcal{N}$ but only find a partial solution.
   Guidelines:
   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: Experimental details are described in detail in the Hyperparameter subsection and the main body of the paper. Our ViT implementations relied on previous, publicly available implementations (Sinkformer [31] and Eureka [25]).
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [No]
   Justification: While the entire development codebase for this project unfortunately cannot be made public at this point, specific parts of the code are available upon justified request.
   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: There is a dedicated section about hyperparameter choices in the appendix.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: All empirical experiments on the QDSFormer were repeatedly performed. Error bars are shown in all plots and standard deviations are given in all tables.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]

Justification: Quantum circuit execution times are explicitly studied. Moreover we provide compute resource details in the hyperparameter section.
Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?
Answer: [Yes]
Justification: Code is respected.
Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: [NA]
Justification: This is a piece of foundational research in quantum machine learning that is currently only applicable to relatively small-scale data (i.e., small images). Beyond the general societal implications of advances in quantum computing hardware, which will be vast and potentially disruptive, certainly for cryptography but potentially also for machine learning, we do not feel that there is anything specific about this paper.
Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
Answer: [NA]
Justification:
Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**
Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [Yes]
Justification: A table with all datasets, citations, size and license terms are explicitly given in appendix. No data is re-distributed, license terms are respected.
Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [NA]
Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [NA]
Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
    Answer: [NA]
    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.
    Answer: [NA]
    Guidelines:
    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.