Quantum DNA Encoder: A Case-study in gRNA Analysis

Nam Nguyen, M.A
Department of Statistics
University of South Florida
Tampa, FL 33620
namphuongnguyen510@gmail.com

***Abstract*—Quantum computing can potentially speed up certain tasks greatly, but its benefits for computational models must be better defined in early literature. This work introduces a new Quantum DNA Encoder (QDE) design that can efficiently and effectively encode genetic data. QDE uses a simple circuit model that can be used on a 4-qubit system and provides better class boundaries than One-Hot Encoding (OHE) by generalizing data encoding to a higher-dimensional embedding space. In order to showcase how well the method works, we utilize a large collection of gRNA within the CRISPR Cas9 system. Comparing baseline representations, QDE outperforms OHE in two out of four representations with a high level of statistical significance. This study is significant in the early quantum machine intelligence literature as the translation of quantum technology to biomedical research. The implementation of our study is available at: https://github.com/namnguyen0510/Quantum-DNA-Encoder.**

***Index Terms*—Quantum Computing, Quantum Biology, DNA Analysis, CRISPR.**

I. Introduction

QUANTUM computing is an emerging discipline that leverages the principles of quantum mechanics for computational purposes. Unlike classical computers that rely on bits representing either 0 or 1, quantum computers employ qubits, which can exist in superposition, simultaneously embodying both 0 and 1. This unique characteristic enables quantum computers to handle and retain extensive information and execute specific computations more effectively than their classical counterparts. Quantum embedding, in the context of quantum computing, refers to a technique used to represent classical data or problems in a quantum format or quantum state. It involves mapping classical information onto a quantum system, such as qubits, to utilize the computational power of quantum algorithms and quantum computers to solve the given problem.

We find that quantum computing could enable better data encoding of the genetic codes. Specifically, the genetic code is spanned by the character set $\mathbb{B} = \{A, T, G, C\}$, and structural information is indispensable to enable learning tasks on DNA codes. However, the dominant approaches for information retrieval of DNA codes on the classical computer do not contain such information. Specifically, mathematical features [7], [8] embeds DNA codes into scalar vector fields, thus lacking geometric information. On the other hand, the conventional encoding for ML models is One-Hot Encoding (OHE), which encodes data into Euclidean vector space, given as

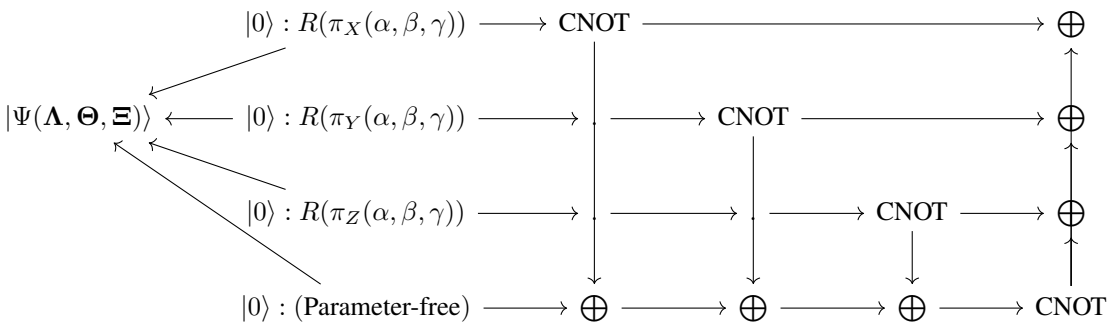
$$\begin{aligned} \mathbb{B} &\to V := \mathbb{R}^4 \\ \boldsymbol{A} &\mapsto [1,0,0,0]^\intercal = \boldsymbol{e}_0 \\ \boldsymbol{T} &\mapsto [0,1,0,0]^\intercal = \boldsymbol{e}_1 \\ \boldsymbol{G} &\mapsto [0,0,1,0]^\intercal = \boldsymbol{e}_2 \\ \boldsymbol{C} &\mapsto [0,0,0,1]^\intercal = \boldsymbol{e}_3 \end{aligned} \tag{1}$$

It is agreeable in the associated literature involving DNA code analysis that structural information matters [10], [11], so geometric information beyond Euclidean must be accounted for in the code encoding. Quantum computing could be a better data encoding approach for the genetic code. Specifically, quantum embedding can be generalized as transforming original data space $\boldsymbol{X}$ on the high-dimensional Hilbert vector space $\mathcal{H}$ of complex numbers through parameterized quantum unitary transformation. Such transformation is considered as model weights in Quantum Neural Networks literature [1], [20]–[23].

Here, we recognize that the geometric interpretation of Pauli-based quantum gates can be used to represent the geometric information within the derived quantum embeddings. In this work, we introduce a novel ansatz design named Quantum DNA Encoder (QDE), which enables efficient and effective data encoding of the genetic codes. First, the model ansatz circuit is low complexity and deployable on a 4-qubit system. Second, the proposed QDE is a generalization of OHE, which encodes data on higher-dimensional embedding space; thus, a better decision boundary could be derived [15]. Finally, we demonstrate our proposed model on a large collection of gRNA in the CRISPR Cas9 system, including $70,892$ gRNAs of length 20 nucleotides (nt). It is worth noting that the demonstrated databases are considerable as large-scale studies in the early quantum machine intelligence literature. two out of four baseline representations from QDE outperformed those from OHE with a high level of statistical significance.

We organize the article as follows:

- **Section** II starts with introducing the QDE ansatz structure (**Section** II-A), followed by a pipeline for genetic dual encoding using the proposed QDE (**Section** II-B).
- **Section** II introduces the background of gRNA analysis in CRISPR technology, followed by experimental design (**Section** III-A). We report the numerical result in **Section** III-B.



- We give a brief literature review and the conclusion in **Section** IV.

II. Methods

A. Quantum DNA Encoder (QDE)

We construct the QDE ansatz circuit by arbitrary single qubit rotation

$$R(\alpha,\beta,\gamma) = R_Z(\gamma)R_Y(\beta)R_Z(\alpha) = \begin{bmatrix} e^{-i(\alpha+\gamma)/2}\cos(\beta/2) & -e^{i} \\ e^{-i(\alpha-\gamma)/2}\sin(\beta/2) & e^{i(\alpha} \end{bmatrix} \tag{2}$$

presented as such unitary transformation over a 4-qubit system

The density matrix corresponding to the electronic wave $|\Psi(\mathbf{\Lambda},\boldsymbol{\alpha},\boldsymbol{\gamma})\rangle$ is given as

$$\mathbf{\Pi} = |\Psi(\mathbf{\Lambda},\mathbf{\Theta},\mathbf{\Xi})\rangle\langle\Psi(\mathbf{\Lambda},\mathbf{\Theta},\mathbf{\Xi})| \tag{3}$$

We physically interpreted the proposed QDE: the transformation on the first three qubits represents the transformation of 3 dimensional space, corresponding to Ox, Oy, and Oz. In contrast, the last parameter-free qubit represents the transformation of time. The time transformation is naturally coherent with space transformation, established by sequential CNOT gates. To form a strong entanglement layout [19], we entangle the last respectively to the first, the second, and the third qubits. We define such entanglement as the feedback of the qubit representing time to qubits representing space.

We use a permutation of the weight set $\{\mathbf{\Lambda},\mathbf{\Theta},\mathbf{\Xi}\}$ to establish the coherency of space transformation while reducing the CNOT gates needed for such entanglement layouts. We have learned a valuable lesson from our previous work on how entanglement affects the accuracy of classifiers. [17]. Specifically, the generalized-unitary transformation on the qubit representing space is performed as in permutation **Table** I and **Table** II.

TABLE I
Homogeneous-Diagonal Permutation (Π^+).

qubit 1	π_X	α	β	γ
qubit 2	π_Y	γ	α	β
qubit 3	π_Z	β	γ	α

We use these two permutations out of $3! = 3 \times 2 \times 1 = 6$ permutations of $\{\alpha,\beta,\gamma\}$ because the first set of weight (Π^+) is created by shifting weights (top row to bottom row) to the right. As a result, we have the same elements in the diagonal (homogeneous-diagonal permutation). Similarly, the second set of weights (Π^-) is created by shifting weights (top row to bottom row) to the left. As a result, we have the same elements in the diagonal (homogeneous, anti-diagonal permutation).

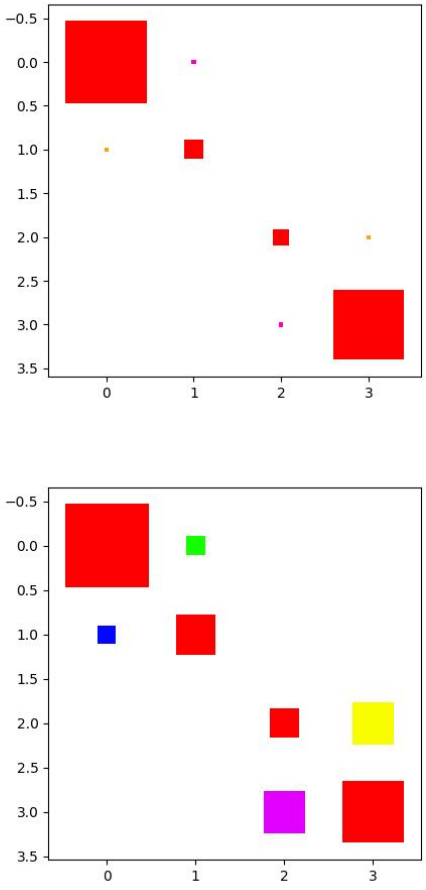


Fig. 1. **The representations, or complex-valued tensors Π^+ (Left) and Π^- (Right) from permutation of weight sets in Table I and Table II.**

We show the variational encoding from QDE, including homogeneous-diagonal (Π^+) and homogeneous, anti-diagonal (Π^-) matrix in **Animation**-S1 and S2. Of note, these matrices have complex number entries. We show the representation, or

TABLE II
HOMOGENEOUS, ANTI-DIAGONAL PERMUTATION (Π^-).

qubit 1	π_X	α	β	γ
qubit 2	π_Y	β	γ	α
qubit 3	π_Z	γ	α	β

complex-valued matrices Π^+ and Π^- in **Figure** 1. We use the package [13] for complex-valued matrices visualization, using Hinton diagram [12]. Specifically, every square represents a matrix element $\boldsymbol{z} = a + bi, (a, b) \in \mathbb{R}^2$, and its size corresponds to the modulus $|\boldsymbol{z}| = \sqrt{a^2 + b^2}$, similar to a Hinton diagram. However, instead of using only black and white to represent positive and negative values, a cyclic color map is used to assign colors to the squares, indicating the complex number's phases [13].

There are three observations from **Figure** 1:

1) The embedding of QDE is a higher-dimensional embedding than Euclidean embeddings since each entry is a complex number $\boldsymbol{z} = a + bi$ formed by the Cartesian product of $\mathbb{R}^2$. Thus, producing the same tensor types ($\Pi^{\pm}$) from classical simulations is insufficient since we need to sample two floating points (a, b). In contrast, directly deriving these tensors from quantum hardware only requires sampling of three floating points (α, β, γ).
2) The embedding is similar to one-hot encoding (OHE) in the Euclidean case as the diagonal entries have the largest amplitude (in OHE, is 1). In the proposed tensor $\Pi^{\pm}$, the diagonal amplitude is saturated to neighboring entries, but the diagonal entries still produce the emphasized signal (strongest amplitude).
3) The QDE has double space complexity compared to OHE. Specifically, the OHE for n-bit DNA codes is of length $n\times[4]$, while QDE needs $n\times[4\times2]$ since storing a complex number requires two floating points. Here, we emphasize the quantum supremacy of using QDE: we only need to sample three floating points for the QDE blocks as shown in **Equation** II-A.

B. Genetic Dual Encoder

We will present the protocol to use QDE for genetic code encryption. Here, we use encryption because we scramble the weight set using the permutation group formed by (α, β, γ) like cryptography techniques. We use both genetic string and its dual partners for encryption. Specifically, given an input genetic string $\boldsymbol{g} = (b_1, b_2, \ldots, b_n)$ of length n (ordered set) for which $b_i \in \mathbb{B} = \{A, T, G, C\}$, the dual genetic string $\bar{\boldsymbol{g}} = (\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_n)$ is constructed as the binding rule

1) A binds T, so if $b_i = A$ then $\bar{b}_i = T$ and if $b_i = T$ then $\bar{b}_i = A$.
2) G binds C, so if $b_i = G$ then $\bar{b}_i = C$ and if $b_i = C$ then $\bar{b}_i = G$.

Thus, we present the encoding pipeline as the following mapping The final embedding is derived by applying a binary relation of the representation $\boldsymbol{H}$. Here, we consider the following aggregation techniques:

TABLE III
DIMENSIONALITY ANALYSIS OF EVALUATED EMBEDDINGS.

Technique	Vector Space	Original Shape	Flattened Shape
OHE	$\mathbb{R}$	$n \times [20 \times 4]$	$n \times 80$
QDE-MA	$\mathbb{C}$	$n \times [20 \times 4]$	$n \times 80$
QDE-CD	$\mathbb{C}$	$n \times [20 \times 8]$	$n \times 160$
QDE-CW	$\mathbb{C}$	$n \times [40 \times 4]$	$n \times 160$
QDE-CO	$\mathbb{C}$	$n \times [20 \times 20]$	$n \times 400$

1) (QDE-MA) Mean Average: $\bar{\boldsymbol{H}} = \frac{1}{2}\Big(\boldsymbol{X}^{\Pi^+}_{n\times4} + \boldsymbol{X}^{\Pi^-}_{n\times4}\Big)_{n\times4}$.
2) (QDE-CD) Depth-wise Concatenate: $\boldsymbol{H}^{\oplus}_{\boldsymbol{D}} = \Big(\boldsymbol{X}^{\Pi^+}_{n\times4} \bigoplus_D \boldsymbol{X}^{\Pi^-}_{n\times4}\Big)_{n\times8}$.
3) (QDE-CW) Breadth-wise Concatenate: $\boldsymbol{H}^{\oplus}_{\boldsymbol{B}} = \Big(\boldsymbol{X}^{\Pi^+}_{n\times4} \bigoplus_B \boldsymbol{X}^{\Pi^-}_{n\times4}\Big)_{2n\times4}$.
4) (QDE-CO) Commutator Operator: $\boldsymbol{H}^{\otimes}_{\boldsymbol{O}} = \Big(\boldsymbol{X}^{\Pi^-}(\boldsymbol{X}^{\Pi^+})^\intercal - \boldsymbol{X}^{\Pi^+}(\boldsymbol{X}^{\Pi^-})^\intercal\Big)_{n\times n}$.

We tested the commutator representation $\boldsymbol{H}^{\otimes}_{\boldsymbol{I}} = \Big((\boldsymbol{X}^{\Pi^-})^\intercal \boldsymbol{X}^{\Pi^+} - (\boldsymbol{X}^{\Pi^+})^\intercal(\boldsymbol{X}^{\Pi^-})\Big)_{4\times4}$; however, the derived features are non-representative across input gRNA due to too small embedded size (4×4, See **SuppMat**-S1).

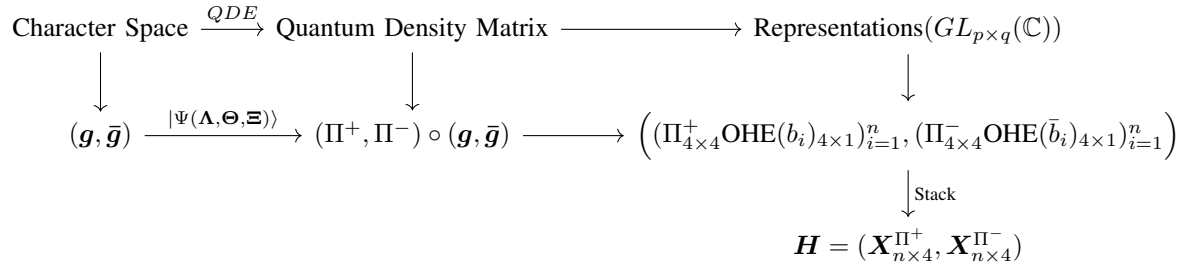
III. RESULTS

A. Case-study

We evaluate the quality of derived embedding sets using the gRNA of the CRISPR-Cas 9 form Achilles project [2]. CRISPR stands for "Clustered Regularly Interspaced Short Palindromic Repeats," which are unique DNA sequences found in the genomes of bacteria and other organisms. Cas9, conversely, refers to a protein called "CRISPR-associated protein 9" that acts as a molecular scissor in the system. The way the CRISPR-Cas9 system works is by using a guide RNA (gRNA) molecule specifically created to target a certain DNA sequence.

1) Dataset: The evaluated dataset includes $70,892$ gRNA of 20 nucleotides (nt). The targeted variable $\boldsymbol{Y}$ is the efficacy score, which is a therapeutic score indicating how effective the gRNA knock-off the correct genes. The range of $\boldsymbol{Y}$ is normalized in $[0, 1]$. We will compare our proposed feature sets, including mean average (QDE-MA), Depth-wise Concatenate (QDE-CD), Breadth-wise Concatenate (QDE-CW), and Commutator Operator (QDE-CO) (**SuppMat**-S2) with the conventionally dominant OHE.

2) Evaluation Metrics: We compare the quality of embedding matrices $\boldsymbol{H}$ by measuring its relevancy [18] with the target variable $\boldsymbol{Y}$. Specifically, for the dataset of n gRNAs, we flatten the embeddings into $1D$ vectors, resulting in embedded matrices $\boldsymbol{H}_{n\times d}$, d is the dimensionality of flattened vectors. The dimensionality analysis of each embedding is given in **Table** III, associated with 20-nt gRNA inputs. Good embeddings



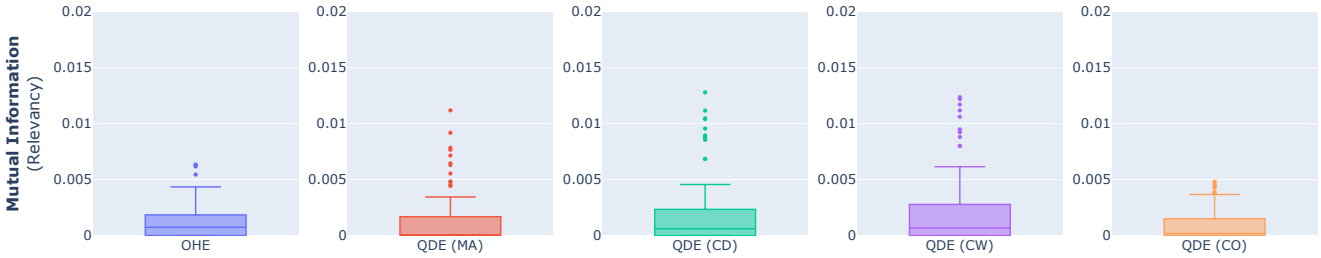


Fig. 2. **The Relevancy of Compared Feature Sets toward The Efficacy Score $\boldsymbol{Y} \in [0, 1]$.**

have higher mutual information toward the target variable $\boldsymbol{Y}$, computed by

$$MI(\boldsymbol{X},\boldsymbol{Y}) = \sum\sum P(x,y)\log\left(\frac{P(x,y)}{P(x)\cdot P(y)}\right) \quad (4)$$

where:

- $P(x, y)$ represents the joint probability mass function (PMF) of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$.
- $P(x)$ and $P(y)$ represent the marginal PMFs of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively.

3) Experimental Environments: All experiments used Python 3.7.0, numpy 1.21.5, sci-kit-learn 1.0.2, and pennylane 0.22.1 on an Intel i9 processor (2.3 GHz, eight cores), 16GB DDR4 memory. The repository for implementation is given in https://github.com/namnguyen0510/Quantum-DNA-Encoder.

B. Effectiveness Analysis

We show the relevancy of evaluated feature sets to efficacy score $\boldsymbol{Y}$ in **Figure** 2. The QDE-MA, QDE-CD, and QDE-CW embeddings outperformed OHE with statistical significance under level $\alpha = 0.05$, showed in **Table** IV. Besides, the QDE-CO is the lowest quality embeddings with the highest dimensionality, meaning the commutative operator is ineffective in deriving a good representation. Thus, two out of four representations constructed by the proposed dual QDE, involving group actions on set $\boldsymbol{H} = (\boldsymbol{X}^{\Pi^+}_{n\times4}, \boldsymbol{X}^{\Pi^-}_{n\times4})$ are effective and outperformed the classical counterpart.

We visualize the complex-valued representations of two input gRNAs, AAAAAAATCCAGCAATGCAG and AAAAAGACAACCTCGCCCTG in **SuppFig** 1 and 2, respectively. Notably, the repetition patterns of A's are well-represented in lower-dimensional embeddings QDE-MA, QDE-CD, and QDE-CW like in OHE (not shown). However, the difference

TABLE IV
STATISTICS TEST TO COMPARE THE QUALITY OF FEATURE SETS. THE STATISTICALLY SIGNIFICANT TESTS ARE HIGHLIGHTED WITH *.

Comparison (One-sided)	Test Statistic (t-stat)	p-value
OHE less than QDE-MA	-0.7038881	0.24136449
OHE less than QDE-CD	-1.77567669	**0.038564***
OHE less than QDE-CW	-2.58523493	**0.00516862***
OHE less than QDE-CO	2.07785468	0.97978906
QDE-MA less than OHE	0.7038881	0.75863551
QDE-MA less than QDE-CD	-0.70259851	0.24166321
QDE-MA less than QDE-CW	-1.39672946	0.08213645
QDE-MA less than QDE-CO	2.16677097	0.98347842
QDE-CD less than OHE	1.77567669	0.961436
QDE-CD less than QDE-MA	0.70259851	0.75833679
QDE-CD less than QDE-CW	-0.85754802	0.19589678
QDE-CD less than QDE-CO	4.13259594	0.99997276
QDE-CW less than OHE	2.58523493	0.99483138
QDE-CW less than QDE-MA	1.39672946	0.91786355
QDE-CW less than QDE-CD	0.85754802	0.80410322
QDE-CW less than QDE-CO	4.91050799	0.99999898
QDE-CO less than OHE	-2.07785468	**0.02021094***
QDE-CO less than QDE-MA	-2.16677097	**0.01652158***
QDE-CO less than QDE-CD	-4.13259594	**2.72365221e-05***
QDE-CO less than QDE-CW	-4.91050799	**1.01776762e-06***

is highlighted in high-dimensional embedding QDE-CO as the repeated patterns are located in different rows and columns of the corresponding representation matrices (bottom-right). The embeddings of the proposed QDE generalize OHE since each complex diagonal entry can intake unit value like in OHE. Besides, we also emphasize the quantum advantages of QDE over the classical OHE in the ending observation of **Section** II-A

IV. CONCLUSION

To this end, we have introduced the QDE and its implementation with a low-complexity ansatz structure on the 4-

qubit quantum system in **Equation** II-A. We then proposed four sets of representations derived from the quantum density matrices parameterized by permutations of three floating points (α, β, γ). Besides, we also highlight the advantages of the proposed encoding methods over the classical OHE. We showed that QDE is a generalization of OHE on complex number vector space, and also, simulating the same complex-valued representations from our QDE is inefficient using classical computing.

We demonstrate the proof of concept using a broad analysis of gRNAs in CRISPR-Cas 9 system. We showed that the proposed QDE is better than OHE, as two out of four derived feature sets outperform the classical counterpart. It is early to claim that our proposed QDE is effective in any learning tasks regarding genetic codes. However, we see two main contributions to our study. First, the proposed DQE is one of the first research to translate the technical advantages of quantum computing [16], [17] to biomedical problems [14]. Second, the proposed framework is a promising approach complementary to current advances in Quantum Machine Learning literature [1], [3]–[6], [9] since any ML model can be applied on feature sets derived from the proposed DQE.

REFERENCES

[1] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021.

[2] A. J. Aguirre, R. M. Meyers, B. A. Weir, F. Vazquez, C.-Z. Zhang, U. Ben-David, A. Cook, G. Ha, W. F. Harrington, M. B. Doshi, et al. Genomic copy number dictates a gene-independent cell response to crispr/cas9 targetinggenomic copy number affects crispr/cas9 screens. *Cancer discovery*, 6(8):914–929, 2016.

[3] M. Benedetti, B. Coyle, M. Fiorentini, M. Lubasch, and M. Rosenkranz. Variational inference with a quantum computer. *Physical Review Applied*, 16(4):044057, 2021.

[4] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019.

[5] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, M. S. Alam, S. Ahmed, J. M. Arrazola, C. Blank, A. Delgado, S. Jahangiri, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.

[6] N. Berner, V. Fortuin, and J. Landman. Quantum bayesian neural networks. *arXiv preprint arXiv:2107.09599*, 2021.

[7] R. P. Bonidia, D. S. Domingues, D. S. Sanches, and A. C. de Carvalho. Mathfeature: feature extraction package for dna, rna and protein sequences based on mathematical descriptors. *Briefings in bioinformatics*, 23(1):bbab434, 2022.

[8] R. P. Bonidia, D. S. Sanches, and A. C. de Carvalho. Mathfeature: feature extraction package for biological sequences based on mathematical descriptors. *bioRxiv*, pages 2020–12, 2020.

[9] S. E. Borujeni, S. Nannapaneni, N. H. Nguyen, E. C. Behrman, and J. E. Steck. Quantum circuit representation of bayesian networks. *Expert Systems with Applications*, 176:114768, 2021.

[10] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[11] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[12] M. Contributors. Hinton diagram demo. https://matplotlib.org/stable/gallery/specialty_plots/hinton_demo.html, Accessed on May 27, 2023. Accessed on May 27, 2023.

[13] J. Gross. Complex matrix visualization. https://jarthurgross.name/2022/04/09/complex-matrix-visualization, April 2022. Accessed on May 27, 2023.

[14] V. Konstantakos, A. Nentidis, A. Krithara, and G. Paliouras. Crispr–cas9 grna efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Research*, 50(7):3616–3637, 2022.

[15] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*, 2020.

[16] N. Nguyen and K.-C. Chen. Bayesian quantum neural networks. *IEEE Access*, 2022.

[17] N. Nguyen and K.-C. Chen. Quantum embedding search for quantum machine learning. *IEEE Access*, 10:41444–41456, 2022.

[18] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[19] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, 2020.

[20] M. Schuld and N. Killoran. Quantum machine learning in feature hilbert spaces. *Physical review letters*, 122(4):040504, 2019.

[21] M. Schuld and N. Killoran. Is quantum advantage the right goal for quantum machine learning? *arXiv preprint arXiv:2203.01340*, 2022.

[22] M. Schuld and F. Petruccione. Quantum models as kernel methods. In *Machine Learning with Quantum Computers*, pages 217–245. Springer, 2021.

[23] M. Schuld, R. Sweke, and J. J. Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021.