用于纵向生物医学研究的量子机器学习框架

Maria Demidik
Deutsches Elektronen-Synchrotron DESY
Zeuthen, Germany
The Cyprus Institute
Nicosia, Cyprus
maria.demidik@desy.de

Filippo Utro
IBM Research
Yorktown Heights,NY, USA
futro@us.ibm.com

Alexey Galda
Moderna
Cambridge, MA, USA
alexey.galda@modernatx.com

Karl Jansen
The Cyprus Institute
Nicosia, Cyprus
Deutsches Elektronen-Synchrotron DESY
Zeuthen, Germany
karl.jansen@desy.de

Daniel Blankenberg
Lerner Research Institute
Cleveland Clinic
Cleveland, Ohio, USA
blanked2@ccf.org

Laxmi Parida
IBM Research
Yorktown Heights, NY, USA
parida@us.ibm.com

摘要—纵向生物医学研究在追踪疾病进展、治疗反应以及耐药机制的出现方面发挥着至关重要的作用，特别是在癌症和神经退行性疾病等复杂疾病的背景下。然而，生物学数据的高度维度与纵向队列规模有限相结合，对传统机器学习方法提出了重大挑战。在这项工作中，我们探讨了量子机器学习（QML）在纵向生物标志物发现中的潜力。我们提出了一种即时量子多项式时间（IQP）特征图的新修改方案，旨在编码多个时间点上的生物医学数据集的时序依赖性。通过对合成和现实世界数据集——包括滤泡淋巴瘤和阿尔茨海默病研究——进行数值模拟，我们展示了我们的纵向 IQP 特征图增强了量子核捕捉单个受试者内部时间模式的能力，为 QML 在临床研究中的应用提供了一个有前景的方向。

***Index Terms*—量子核，纵向分析，滤泡性淋巴瘤，阿尔茨海默病。**

I. 介绍

纵向研究在生物医学研究中对于捕捉时间跨度上的生物学、生理学和认知过程的动态变化至关重要。与仅提供快照的横断面研究不同，纵向数据能够追踪个体轨迹，识别疾病进展的早期标志，并评估治疗效果或抗性机制 [1] 。这些见解在慢性及进行性疾病中尤为重要。例如，许多癌症疗法最初抑制肿瘤生长，但最终由于抗性机制的出现而失效。在滤泡性淋巴瘤（FL）的情况下，转化为更具侵袭性的疾病状态仍然是一个关键的临床挑战 [2] 。类似地，阿尔茨海默病（AD）以渐进神经退化过程为特征，在患者认知功能长时间看似稳定后会出现快速衰退 [3] 。

纵向数据由同一组受试者在不同时间点重复测量的一系列特征组成，这使得可以研究受试者的个体变异性和时间趋势。尽管纵向生物医学数据具有临床相关性，但这些数据通常会受到样本量小、采样间隔不规则或稀疏以及高昂的采集成本的影响，特别是涉及组学或成像模式时。这些因素对传统机器学习方法构成了重大挑战，因为这些方法通常需要大量的标注数据才能有效泛化。

量子机器学习（QML）为经典方法提供了一种有前景的替代方案 [4]，特别是在样本量有限的情况下。值得注意的是，QML 模型已经显示可以从较少的数据点中进行泛化 [5]，并通过量子核提供了数据依赖性的预测优势 [6]。这些特性使得 QML 成为纵向研究的有力工具 [7]，并且在更广泛的生物医学和药物发现领域也得到了探索 [8]。

然而，大多数现有的量子机器学习模型都是在输入数据独立同分布（i.i.d.）的假设下开发的，这一假设未

能考虑到纵向数据中存在的时间依赖性。在生物医学应用中，忽略这种结构可能导致表示次优和预测性能降低。

在本研究中，我们提出了一种专门针对纵向生物医学研究的量子核框架。已有研究表明，将核函数定制到目标应用可以增强 QML 模型的性能 [9]。在量子核方法中，经典数据通过特征映射编码为量子态，这定义了核的结构。一种广泛采用的特征映射是瞬时量子多项式时间（IQP）特征映射 [10]。我们引入了一种纵向 IQP 特征映射，该映射明确地在多个时间点之间纳入了时间依赖性，从而扩展了量子核对时间序列生物医学数据的表示能力。

我们对提出的特征图在合成数据和两个公共纵向生物医学数据集上进行了评估。结果显示，与 IQP 特征图相比，纵向 IQP 特征图能更好地建模疾病动态。

本文的其余部分结构如下。第 II节介绍核方法所需的背景知识。第 III节介绍了 IQP 和纵向 IQP 特征图，以及说明它们建模时间依赖性的数值模拟。第 IV节描述了在真实世界生物医学数据集上的经验评估。最后，第 V节讨论了我们发现对纵向生物标志物发现和更广泛的生物医学应用的影响。

II. 框架

支持向量机（SVMs）为监督学习中的分类提供了一个成熟的框架 [11] 。通过利用高维特征映射，SVMs 将输入数据投影到转换后的特征空间，在该空间中构造一个最优超平面以分离不同的类别。这种方法依赖于核技巧，它可以通过核函数高效地计算高维空间中的内积 [12] 。

设 $\boldsymbol{x}$ 和 $\boldsymbol{x}'$ 是两个长度为 n 的实值输入向量（n 是特征的数量），并令 $\phi(\boldsymbol{x})$ 表示相关的特征图。核技巧允许在特征空间中计算内积而不显式评估 ϕ，使用一个核函数 K 使得

$$K(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle. \tag{1}$$

特征映射 ϕ 的选择（由 K 隐式定义）在模型捕捉复杂非线性决策边界的能力中起着关键作用。一个广泛使用的例子是径向基函数（RBF）核，定义为

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma ||\boldsymbol{x} - \boldsymbol{x}'||^2), \tag{2}$$

其中 $\gamma > 0$ 是控制核对距离敏感性的比例参数。

量子计算机通过参数化量子电路提供了一种定义特征映射的替代框架。令 $|0\rangle^{\otimes n}$ 表示 n-量子比特计算基态。一个作用于 n 个量子比特的参数化幺正操作 $U(\boldsymbol{x})$，通过将经典输入 $\boldsymbol{x}$ 编码到量子态 $|\psi(\boldsymbol{x})\rangle = U(\boldsymbol{x})|0\rangle^{\otimes n}$ 中，定义了一个量子特征映射。相应的核函数 K 是通过量子态之间的平方重叠（保真度）构建的，

$$K(\boldsymbol{x}, \boldsymbol{x}') = |\langle \psi(\boldsymbol{x}) | \psi(\boldsymbol{x}') \rangle|^2 , \tag{3}$$

可以通过图 1所示的量子电路进行计算。

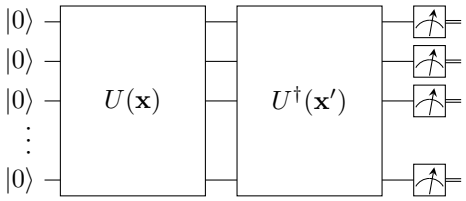

图 1: 用于评估量子核的量子电路。核 $K(\boldsymbol{x}, \boldsymbol{x}')$ 是通过测量两个量子态之间的平方重叠来计算的，定义见公式 (3)。

在本研究中，我们采用了一种基于瞬时量子多项式时间（IQP）电路的量子特征映射 [13] ，这是一类在量子复杂性理论中扮演重要角色的量子电路，并且广泛用作量子机器学习中的核函数 [10], [14] 。

III. 纵向数据嵌入

形式上，一个 IQP 特征映射将长度为 n 的输入 $\boldsymbol{x}$ 编码到量子态中作为

$$|\psi(\boldsymbol{x})\rangle = U_D(\boldsymbol{x}) H^{\otimes n} |0\rangle^{\otimes n}, \tag{4}$$

其中 U_D 是一个对角酉算子，通常由参数化的单比特和双比特旋转组成，如 R_Z 和 R_{ZZ}。

最直接的通过 IQP 特征映射编码时间数据的方法是将每个时间点独立地作为特征映射的层进行编码。然后，输入 $\boldsymbol{x}$ 由 T 个时间点表征到量子态中的编码给定为

$$|\psi(x)\rangle = U_D\left(\boldsymbol{x}^T\right) H^{\otimes n} \cdots U_D\left(\boldsymbol{x}^1\right) H^{\otimes n} |0\rangle^{\otimes n}, \tag{5}$$

其中每个 $U_D\left(x^t\right)$ 都由时间点 t 处的输入参数化。图 2 提供了对用于编码纵向数据的 IQP 特征映射的说明。

虽然这种方法能够使用 IQP 特征映射对时间结构化的输入进行编码，但它将每个时间点视为独立的层，因此无法捕捉单一样本内部的时间相关性。这可能导致

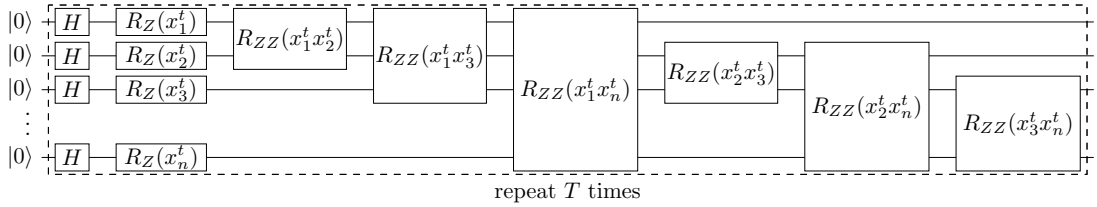

图 2: 用于定义纵向数据量子特征映射的 IQP 电路。输入由 n 个特征组成，这些特征在 T 个时间点上被测量，代表了纵向观测值。对于每个时间点 $t \in \{1, \ldots, T\}$，应用一个独立的 IQP 层来编码相应的特征向量 $\boldsymbol{x}^t$。每层包括 Hadamard 门，随后是一个由单量子比特和双量子比特旋转（如 R_Z）组成的对角单位元 $U_D(\boldsymbol{x}^t)$。R_{ZZ}

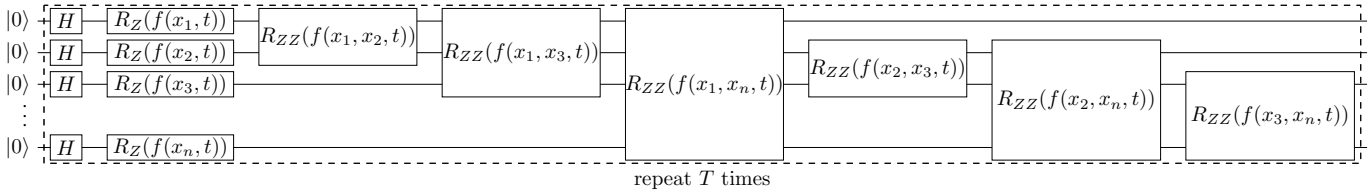

图 3: 纵向 IQP 电路用于编码时间依赖的特征交互。与为每个时间点应用单独一层的 IQP 特征映射不同，纵向 IQP 电路通过在每一层评估的时间依赖函数来编码数据。具体来说，量子电路参数由形式为 $f(x_i, t)$ 和 $f(x_i, x_j, t)$ 的函数调制，分别如等式 (6) 和 (7) 中定义的那样。这些函数捕捉了时间动态和成对特征交互，使得纵向数据能够更丰富地嵌入到量子状态空间中。

缺乏样本内时间相关性的建模。为了解决这一限制并考虑样本内的时间依赖性，我们建议通过对如何利用输入特性来参数化量子门进行修改来改进 IQP 特征映射。而不是将每个时间点独立处理，我们在特征编码中定义了一个随时间变化的功能依赖关系。具体来说，对于层 t（对应于 t 时间点），单量子比特门的输入参数如下获得：

$$f(x_i, t) = \frac{1}{t} \sum_{m=1}^{t} x_i^m \,, \tag{6}$$

其中 x_i^m 表示特征 i 在时间点 m 的值。相应地，第 t 层的两量子比特门的参数由以下给出：

$$f(x_i, x_j, t) = f(x_i, t) \cdot f(x_j, t). \tag{7}$$

这种累积形式引入了时间平滑效果，使得核函数能够捕捉到随着时间推移的渐进性个体内变化。带有修改编码的 IQP 特征映射被称为纵向 IQP 特征映射，并在图 3 中进行了说明。

我们使用合成数据集评估 IQP 和纵向 IQP 特征图。生成的数据集包含 100,000 个样本，每个样本由一个取值范围在 [0,2π] 的单一特征描述。为了模拟时间结构，每个样本包括两个时间点，并通过计算每个样本与固定参考样本（两个时间点的值都设为 $\pi/2$）之间的保真度来建模依赖关系。

如图 4a所示，对于 IQP 特征映射，在第一个时间点固定的情况下，生成样本与平稳样本之间的保真度会随着第二个时间点的变化而周期性和对称性变化。这种在单一样本内失去的时间依赖性对于生物医学应用是不希望看到的。因此，在纵向研究中利用 IQP 特征映射可能会导致 QML 模型的表现有限。

相比之下，在图 4b中，纵向 IQP 特征图使核函数能够捕捉第二个时间点相对于第一个时间点的变化。这种行为表明纵向 IQP 特征图有可能有效建模时间模式，这对于涉及疾病进展或治疗反应的生物医学研究尤其相关。

IV. 数值结果

为了评估我们的方法，我们使用了与以下疾病相关的两个公开生物医学数据集：

a) 滤泡性淋巴瘤（FL）： 是最常见的惰性 B 细胞非霍奇金淋巴瘤类型。在参考文献 [15] 中，作者进行了一项多组学分析，以研究早期复发和滤泡性淋巴瘤向侵袭性疾病转化的相关情况，这与不良预后相关联。尽管在滤泡性淋巴瘤发病机制中识别了新的生物标志

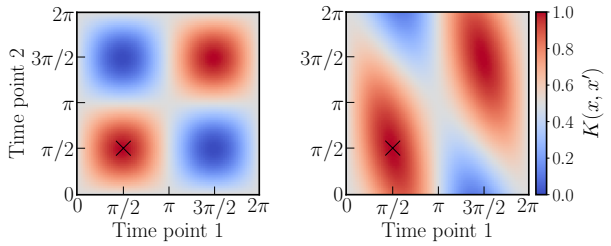

图 4: IQP 和纵向 IQP 核在两时间点单特征设置下的比较。核是相对于固定参考点 $(\pi/2, \pi/2)$ 计算的，并在整个可能的 $x^1, x^2 \in [0, 2\pi]$ 值范围内进行评估，分别对应于时间和 2。左图（a）显示了使用 IQP 特征映射获得的核值，而右图（b）则展示了纵向 IQP 特征映射的结果。参考点 $(\pi/2, \pi/2)$ 在两个图中都用一个 'X' 标记。

物，但仍然难以识别那些会发生所建议的生物标志物突变的个体。基于来自参考文献 [15] 的可用 RNA-seq 数据，我们考虑将 32 名受试者分类为两组的任务：复发未转化滤泡性淋巴瘤（nFL）和转化滤泡性淋巴瘤（tFL）。每位受试者通过多次活检进行表征，数量从一到三次不等。我们在运行核方法之前利用 Welch's t 检验选择了前 20 个显著差异表达的基因。

b) 阿尔茨海默病 (AD)： 是一种进行性的神经退行性疾病。数据由开放访问系列影像研究（OASIS）[3] 收集。150 名受试者的数据显示了 3 或 4 次个体 MRI 扫描的纵向收集情况。此外，受试者的神经退行性状态通过临床痴呆评定量表（CDR）、标准化全脑体积（nWBV）、估计颅内总体积（eTIV）、简易精神状态检查分数（MMSE）和图谱缩放因子（ASF）来描述。最初 64 名受试者被诊断为痴呆，而 72 名受试者在整个研究期间保持非痴呆组。另有 14 名受试者的状况从非痴呆转变为痴呆。遵循 [16] 的参考方法，我们利用神经退行性状态的描述特征基于 CDR 状态进行受试者的二元分类。未来的工作可以通过将 MRI 扫描数据纳入框架来扩展。

我们评估了三种特征图的表现：经典 RBF、IQP 和纵向 IQP，针对两个生物医学纵向数据集。为了将 RBF 核应用于纵向数据，我们将所有的时间点作为独立的特征提供。使用这些特征图，我们通过标准凸优化技术 [16] 训练 SVM 模型。我们在表 I 中报告了二分类的训练和测试准确率。

表 I: 不同特征图在数据集上分类性能的比较。表格报告了使用三种类型的特征图（经典 RBF、IQP 和纵向 IQP）对两个数据集——滤泡性淋巴瘤和阿尔茨海默病的训练和测试准确率。

数据集	特征图	训练精度	测试精度
Follicular lymphoma	RBF	25 / 25	6 / 7
	IQP	25 / 25	6 / 7
	Longitudinal IQP	25 / 25	7 / 7
Alzheimer's disease	RBF	120 / 120	19 / 30
	IQP	111 / 120	18 / 30
	Longitudinal IQP	110 / 120	22 / 30

尽管所有模型都达到了完美的或接近完美的训练准确率，纵向 IQP 特征图在测试性能上相对于 IQP 方法始终表现出改善。这表明了增强的泛化能力，特别是在建模具有潜在时间性或渐进结构的数据时。在两个数据集上，结果表明通过纵向 IQP 特征图纳入时间依赖性对分析如 AD 和 FL 这类渐进动态疾病是有益的。

V. 讨论

在这项工作中，我们介绍了纵向 IQP 特征映射，旨在编码纵向生物医学数据集中的内在时间依赖性。所提出的特征映射明确地包含了样本内部各时间点之间的时间相关性。纵向 IQP 特征映射基于使用累积编码方案的 IQP 特征映射：每个电路层汇集当前和所有先前时间点的信息。这使得量子核能够捕捉到个体内的时序相关性，这对于理解疾病进展的动力学至关重要。

数值实验在生物医学数据集上表明，纵向 IQP 特征图增强了对时间结构化数据的核表达能力。特别地，我们观察到，虽然 RBF 核和基于 IQP 特征图的核实现了高训练精度，但利用纵向 IQP 特征图始终导致更高的测试性能。这些发现强调了将量子特征图适应以反映输入数据结构的价值，特别是在需要关键建模时间进展的背景下。

纵向生物医学数据集通常由小队列组成。虽然量子机器学习模型在低数据环境下具有良好的泛化性能潜力，但仍难以找到一个足够大的公开数据集来评估所提模型的泛化能力。未来的工作应评估复杂时间模式下更大队列中的纵向 IQP 特征图的稳健性和可扩展性。此外，将这种方法扩展到多模态数据，例如结合纵向成像与分子谱，可以进一步增强其在神经退行性疾病和肿瘤疾病生物标志物发现方面的实用性。探索替代量子机器

学习架构，如量子水库计算 [17] 或量子递归模型 [18]，也可能为序列生物医学任务提供互补优势。

最后，我们注意到所有特征图的结果，包括 RBF，可能受益于额外的超参数调整。然而，这项工作的主要目标是强调将量子机器学习模型适应特定领域数据结构的重要性。即使短期内的量子模型仍然可以被经典模拟，它们的设计和评估为疾病动态建模和推进数据驱动的临床研究提供了新颖的计算视角。

致谢

此项工作得到了勃兰登堡州科学、研究和文化部在量子技术与应用中心（CQTA）框架内的资金支持。本项目在 QUEST 框架内，由欧盟地平线欧洲计划（HORIZON）通过 ERA Chair 方案资助，资助协议编号为 101087126。

参考文献

[1] Gad Getz, Carrie Cibulskis, Ignaty Leshchiner, Megan Hanna, Dimitri Livitz, Kara Slowik, Chaya Levovitz, Filippo Utro, Kahn Rhrissorrakrai, Denisse Rotem, Gregory Gydush, Sarah C. Reed, Justin Rhoades, Gavin Ha, Samuel S. Freeman, Christopher Lo, Mark Fleharty, Justin Abreu, Katie Larkin, Michelle Cipicchio, Brendan Blumenstiel, Matt DeFelice, Jonna Grimsby, Susanna Hamilton, Niall Lennon, Viktor A. Adalsteinsson, and Laxmi Parida. Abstract 3001: Broad/IBM Project: Discovery of treatment resistance mechanisms through use of liquid biopsy genomics services. *Cancer Research*, 78(13 Supplement):3001–3001, 07 2018.

[2] Baoyan Bai, Jillian F Wise, Daniel Vodák, Sigve Nakken, Ankush Sharma, Yngvild Nuvin Blaker, Marianne Brodtkorb, Vera Hilden, Gunhild Trøen, Weicheng Ren, et al. Multi-omics profiling of longitudinal samples reveals early genomic changes in follicular lymphoma. *Blood Cancer Journal*, 14(1):147, 2024.

[3] Daniel S. Marcus, Anthony F. Fotenos, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, December 2010.

[4] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

[5] Matthias C. Caro, Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles. Generalization in quantum machine learning from few training data. *Nature Communications*, 13(1):4919, August 2022.

[6] 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, May 2021.

[7] Frederik F. Flöther, Daniel Blankenberg, Maria Demidik, Karl Jansen, Raga Krishnakumar, Rajiv Krishnakumar, Nouamane Laanait, Laxmi Parida, Carl Saab, and Filippo Utro. How quantum computing can enhance biomarker discovery for multi-factorial diseases, December 2024.

[8] Anthony M Smaldone, Yu Shee, Gregory W Kyro, Chuzhi Xu, Nam P Vu, Rishab Dutta, Marwa H Farag, Alexey Galda, Sandeep Kumar, Elica Kyoseva, et al. Quantum machine learning in drug discovery: Applications in academia and pharmaceutical industries. *arXiv preprint arXiv:2409.15645*, 2024.

[9] Jennifer R. Glick, Tanvi P. Gujarati, Antonio D. Córcoles, Youngseok Kim, Abhinav Kandala, Jay M. Gambetta, and Kristan Temme. Covariant quantum kernels for data with group structure. *Nature Physics*, 20(3):479–483, March 2024.

[10] 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.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273 – 297, September 1995.

[12] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220, 2008.

[13] Michael J. Bremner, Ashley Montanaro, and Dan J. Shepherd. Average-case complexity versus approximate simulation of commuting quantum computations. *Phys. Rev. Lett.*, 117:080501, Aug 2016.

[14] Ruslan Shaydulin and Stefan M. Wild. Importance of kernel bandwidth in quantum machine learning. *Phys. Rev. A*, 106:042407, Oct 2022.

[15] Baoyan Bai, Jillian F Wise, Daniel Vodák, Sigve Nakken, Ankush Sharma, Yngvild Nuvin Blaker, Marianne Brodtkorb, Vera Hilden, Gunhild Trøen, Weicheng Ren, Susanne Lorenz, Michael S Lawrence, Ola Myklebost, Eva Kimby, Qiang Pan-Hammarström, Chloé B Steen, Leonardo A Meza-Zepeda, Klaus Beiske, Erlend B Smeland, Eivind Hovig, Ole Christian Lingjærde, Harald Holte, and June Helen Myklebust. Multi-omics profiling of longitudinal samples reveals early genomic changes in follicular lymphoma. *Blood Cancer J.*, 14(1):147, August 2024.

[16] Gopi Battineni, Nalini Chintalapudi, and Francesco Amenta. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (svm). *Informatics in Medicine Unlocked*, 16:100200, 2019.

[17] Yudai Suzuki, Qi Gao, Ken C. Pradel, Kenji Yasuoka, and Naoki Yamamoto. Natural quantum reservoir computing for temporal information processing. *Scientific Reports*, 12(1):1353, January 2022.

[18] Johannes Bausch. Recurrent Quantum Neural Networks, June 2020.