IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 045076 https://doi.org/10.1088/2632-2153/ae304f

MACHINE LEARNING
Science and Technology

PAPER

Encoding molecular structures in quantum machine learning

Choy Boy[1,2,*], **Edoardo Altamura**[1,2,4], **Dilhan Manawadu**[2], **Ivano Tavernelli**[3], **Stefano Mensa**[2] and **David J Wales**[1]

[1] Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
[2] The Hartree Centre, STFC, Sci-Tech Daresbury, Warrington WA4 4AD, United Kingdom
[3] IBM Quantum, IBM Research Zurich, Rüschlikon 8803, Switzerland
[4] Shared first-authorship.
[*] Author to whom any correspondence should be addressed.

E-mail: bc537@cam.ac.uk

Keywords: quantum computing, computational chemistry, emerging technologies

OPEN ACCESS

RECEIVED
27 July 2025

REVISED
20 November 2025

ACCEPTED FOR PUBLICATION
22 December 2025

PUBLISHED
31 December 2025

Abstract

Quantum machine learning (QML) has great potential for the analysis of chemical datasets. However, conventional quantum data-encoding schemes, such as fingerprint encoding, are generally unfeasible for the accurate representation of chemical moieties in such datasets. In this contribution, we introduce the quantum molecular structure encoding (QMSE) scheme, which encodes the molecular bond orders and interatomic couplings expressed as a hybrid Coulomb-adjacency matrix, directly as one- and two-qubit rotations within parametrised circuits. We show that this strategy provides an efficient and interpretable method in improving state separability between encoded molecules compared to other fingerprint encoding methods, which is especially crucial for the success in preparing feature maps in QML workflows. To benchmark our method, we train a parametrised ansatz on molecular datasets to perform classification of state phases and regression on boiling points, demonstrating the competitive trainability and generalisation capabilities of QMSE. We further prove a fidelity-preserving chain-contraction theorem that reuses common substructures to cut qubit counts, with an application to long-chain fatty acids. We expect this scalable and interpretable encoding framework to greatly pave the way for practical QML applications of molecular datasets.

1. Introduction

The integration of machine learning (ML) techniques in chemistry has led to significant advances, such as improved prediction of protein structures [1] and the estimation of blood-brain barrier permeability for small molecules as potential drug candidates [2]. Quantum computing is a promising approach for enhancing ML pipelines involving classical and quantum data [3]. In this context, quantum machine learning (QML) algorithms have been proposed to improve *in-silico* screening and quantum-assisted drug design [4–6]. In the near term, QML may continue to deliver practical advantages in specialised tasks, such as learning from quantum data, simulating physical systems, and employing quantum feature spaces, particularly when paired with hybrid quantum–classical architectures. As quantum hardware is set to improve with longer qubit coherence times [7–9], reduced leakage [10], and suppressed cross-talk [11], QML models may outperform their classical counterparts in representing and optimising high-dimensional, structured, and entangled data, especially in domains like quantum chemistry, material science, and drug discovery. This potential is expected to become even more significant in the fault-tolerant quantum computing (FTQC) regime, where QML is expected to offer competitive speedups for variational [12] and kernel methods, feature selection [13], and generative models [14, 15]. In addition, various studies involving QML have found significant advantages via the combination of superior generalisability [16] alongside higher accuracies for less training data inventories [17] compared to their classical equivalent. As both software and hardware frameworks continue to advance, QML is

poised to become a foundational element in achieving quantum advantage across computational learning and scientific discovery [18]; even more so when next-generation fault-tolerant quantum devices and algorithms become available.

Besides trainable architectures, a key component of any ML pipeline is the choice of encoding data as input vectors. In particular, molecular representation learning seeks to optimise the transformation of molecular structures as suitable input vectors in trainable models [19]. Standard techniques, such as one-hot encoding [20] and its embedded variants [21] are effective in partially alleviating the 'curse of dimensionality' associated with representing molecular structures by compressing high-dimensional binary vectors into lower-dimensional real-valued representations. In addition, more sophisticated techniques such as graph-based molecular representation learning methods (e.g. group graphs [22]) improve upon atom-level encodings by representing substructures as nodes and encoding connectivity via edges.

Quantum encoding schemes, such as basis encoding, angle encoding, and amplitude encoding, map classical features into data-encoding quantum circuits for QML processes [23]. Basis encoding typically represents a binary molecular fingerprint of length τ directly into τ qubits, but this procedure becomes unfeasible for larger fingerprints. Amplitude encoding reduces qubit requirements to $O(\log\tau)$ by mapping a normalised feature vector into the amplitudes of a quantum state, but preparing arbitrary amplitude-encoded states requires an exponential scaling of two-qubit gates in the worst case [24–26], rendering it impractical on near-term devices. Angle encoding, which parametrises one-qubit rotations with feature values, offers a hardware-efficient alternative, but can suffer from poor state separation and trainability issues, especially when paired with dimension-reduced features to reduce quantum hardware requirements. These limitations motivate development of new encoding techniques that strike a balance between expressivity, trainability, and resource efficiency.

There is growing traction in the number of strategies devised for the deployment of quantum neural networks (QNNs) in predicting chemical phenomena, in particular the training and prediction of potential energy surfaces and molecular force fields [27–29]. Recent studies have also proposed techniques to encode molecular properties in Hilbert spaces accessible via quantum computing. Boiko *et al* [30] introduced stereoelectronics-infused molecular graphs, which enrich traditional molecular graphs by incorporating orbital-centric nodes (e.g. σ, π, σ^*, π^*, lone pairs) and quantified donor–acceptor interactions derived from Natural Bond Orbital analysis [31]. A surrogate graph neural network is trained to predict these features directly from 3D molecular geometries, enabling fast and accurate inference for downstream property prediction. This approach enhances model interpretability and generalises to large biomolecules. Compared to classical descriptors, namely Coulomb matrices [32], SOAP [33], or graph-based encodings such as `ChemProp` [34], surrogate graph neural networks offer superior chemical fidelity by encoding quantum interactions explicitly. The advantages of such models include interpretability and high performance in message-passing neural networks, while their limitations include the initial computational overhead of quantum chemical calculations and the need for dataset-specific retraining when the datasets are extended.

Torabian and Krems [35] proposed a novel isomorphism between quantum circuits and polyatomic molecules, enabling the mapping of circuit architectures to molecular descriptors, such as Coulomb matrices, molecular fingerprints, and Gershgorin discs. These descriptors can be used to predict the performance of quantum support vector machines (QSVMs), offering a strategy to reduce the search space in circuit design. Their method complements efforts to mitigate barren plateaux [36], exponential kernel concentration [37], and noise-induced degradation in kernel methods [2]. They also relate to advances in covariant [38] and Fisher kernels that aim to preserve relevant data structure. The main advantage is the physically interpretable restriction of circuit composition using descriptors well-established in chemoinformatics. However, limitations include the potential ambiguity in reverse-mapping molecules back to unique quantum circuits and scalability concerns for deeply layered architectures or high-qubit-count regimes.

Finally, Kamata *et al* [39] developed the molecular quantum transformer (MQT), a hybrid classical–quantum architecture that uses quantum self-attention to represent and predict molecular ground-state energies. The model encodes bond length–dependent molecular Hamiltonians via parametrised quantum circuits and exploits training on multiple geometries for efficient learning of potential energy surfaces. In contrast to methods like variational quantum eigensolvers (VQE) [40] or meta-VQE [41], which require separate circuit evaluations for each molecular configuration, MQT offers a more data-efficient alternative. It also outperforms classical Transformer models when learning from small datasets and supports pretraining and fine-tuning workflows. However, its reliance on amplitude encoding and large circuit ansätze may limit feasibility on near-term hardware. The work aligns with recent proposals for quantum-enhanced transformers [42] and builds on advances in neural-network quantum states and denoising [43].

Despite early theoretical and experimental advances [44], realising quantum advantage in supervised QML remains challenging in training strategies like QNNs due to barren plateaux, high circuit depth, noise, and the expressivity of parametrised quantum circuits [36, 37, 44–46]. Other works have shown that trainable circuit architectures, such as convolutional QNNs, can be classically simulated [47], attenuating potential advantages of using quantum devices as opposed to classical computers. In some cases, such prospects for exponential speedups over classical ML are found to be generated by explicit or implicit assumptions introduced in mathematical proofs [48], making practical quantum advantage uncertain [49].

In this work, we introduce the *quantum molecular structure encoding* (QMSE) scheme, which explicitly encodes molecular bond orders and interatomic couplings via a hybrid Coulomb-adjacency matrix as parametrised one- and two-qubit rotation gates in the data-encoding quantum circuit in QML workflows. This approach addresses several key challenges identified in recent QML literature. First, rigorous quantum speed-up results for supervised learning tasks suggest that specialised feature maps can yield provable advantages [18, 44], but only if they produce sufficiently distinct quantum states; QMSE's graph-based representation can achieve a broader distribution of fidelities compared to conventional fingerprint (angle) encoding. Second, subtleties in trainability and barren plateau effects have been shown to impede variational QML models [45]; by constructing an encoding that exploits commutativity of two-qubit interactions (e.g. R_{xx} rotations), QMSE provides a more robust optimisation landscape. Third, exponential concentration in quantum kernel methods can render quantum-enhanced similarity measures ineffective for high-dimensional classical data [37]; the one-to-one physical mapping of the QMSE scheme allows for a more bespoke representation of molecular datasets as data-encoding circuits that better reflect chemical similarity, and thus we expect QMSE to alleviate the saturation issues associated with traditional fingerprint-based kernels. Finally, the burgeoning success of classical large language models in few-shot learning [50] underscores the importance of scalable, data-efficient architectures. QMSE draws inspiration from this approach by encoding molecular graphs in a structured, modular fashion, thereby facilitating generalisation in small-dataset regimes typical of chemical screening.

Compared to graph-based classical molecular representation learning, QMSE directly incorporates quantum-chemical insights, such as bond orders, interatomic couplings, and stereochemistry, into the single and entangling quantum gates of circuits in the data-encoding layer of QML, akin to a quantum approximate optimisation algorithm (QAOA) circuit for connected graphs. We demonstrate that QMSE not only reduces resource demands on near-term quantum hardware, but also yields significantly higher training and test accuracies, outperforming standard fingerprint encoding in both classification and regression tasks on chemical datasets. Furthermore, we prove a fidelity-preserving chain-contraction theorem that eliminates common molecular fragments in reducing qubit counts, paving the way for scalable QML applications to long-chain molecules and large datasets.

This paper is organised as follows. In section 2, we review conventional feature encoding schemes and their limitations in QML tasks of molecular datasets. Section 3 describes the QMSE approach by defining the hybrid Coulomb-adjacency matrix and its representation as graph states in quantum circuits. Section 4 describes the chemical datasets used for benchmarking, and section 5 reports numerical results for classification and regression tasks, highlighting improvements in trainability and generalisation compared to equivalent models using features as inputs. In section 6, we discuss the implications of structure encoding in light of recent theoretical findings. Finally, section 7 summarises our contributions and outlines future directions, including extensions to quantum kernel methods and expectations for molecular representations in the early quantum fault-tolerant computing regime.

2. Fingerprint encoding

In supervised QML pipelines, given a dataset $D = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^M, y^M)\}$ with M pairs of input vectors $\mathbf{x}$ and their corresponding output values y, the first step involves the encoding of the input data as initial states in a quantum circuit with N qubits via a chosen feature map. If a variational QML workflow is utilised, as we will employ in this work, a parametrised ansatz circuit with unitary $\hat{U}(\boldsymbol{\theta})$ is subsequently applied to the quantum circuit, and the expectation values of a specified Hamiltonian operator $\hat{H}$ from each datapoint are obtained to evaluate a given loss function to be minimised classically. The regime will then suggest new parameters to be fed back into the ansatz, and the cycle between quantum and classical interfaces repeats until either a maximum number of iterations or a suitable convergence criterion is reached [51]. In the context of chemical datasets, where the input molecules in D may be represented in the form of Simplified Molecular Input Line Entry System (SMILES) strings, the chemical moieties and structural features of each input molecule can be further encoded as classical molecular fingerprints, such as RDKit topological fingerprints [52].

At first glance it may seem natural to employ basis encoding to map the binary sequences of classical molecular fingerprints, i.e. representing a given data point $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_\tau)^{\mathrm{T}}$ of fingerprint length τ as a basis-encoded quantum state, i.e. $|\psi_{BA}\rangle = |x_1 \ldots x_i \ldots x_\tau\rangle$, in the computational basis within the Hilbert space $\mathcal{H}_N \cong (\mathbb{C}^2)^{\otimes N}$, and applying Pauli-$X$ gates to the corresponding qubits of the fiducial state $|0\rangle^{\otimes N}$ for $x_i = 1$ [53]. However, in practice, this procedure requires mapping the number of qubits N linearly to the length of each molecular fingerprint, which defaults to $\tau = 2048$, thus making such a scheme unfeasible. Additionally, although the encoded data points can be expressed efficiently with the lowest possible quantum circuit depth of 1, they only represent a tiny fraction of the total possible number of quantum states within $\mathcal{H}$ with no state overlap. Finding an ansatz that minimises the loss function of the QML task would therefore be an especially challenging endeavour.

One may instead consider using amplitude encoding to represent the normalised classical molecular fingerprint as a quantum state $|\psi_{\mathrm{AM}}\rangle$ in $\mathcal{H}$, thereby drastically reducing the number of qubits required to encode the state to $N = \log_2 \tau$, or $N = 11$ for a default molecular fingerprint. Amplitude encoding also ensures a consistent means of comparing molecules based on the presence or absence of certain chemical moieties, where similar molecular wavefunctions lie in close proximity for $\mathcal{H}$ (and vice versa), thus facilitating the QML task. However, a serious drawback of amplitude encoding is the potential complexity of decomposing the unitary operator $\hat{U}_{\mathrm{AM}}$ that evolves the fiduciary state into $|\psi_{\mathrm{AM}}\rangle$ in terms of its basis one- and two-qubit quantum gates. Although various quantum gate decomposition schemes have been proposed to prepare any arbitrary quantum state [24–26], such formulations generally require an exponentially increasing number of entangling CNOT gates with N, thus making the expression of amplitude-encoded quantum circuits particularly unsuitable for near-term devices.

To alleviate the drawbacks associated with either basis or amplitude encoding, angle encoding may be employed as an efficient representation of classical data as angular amplitudes for rotational gates in the data-encoding circuit [54]. As the default τ is typically too large to implement angle rotation directly, the number of input features in $\mathbf{x}$ can first be decreased via standard dimensionality techniques, such as principal component analysis (PCA) to its compressed counterpart $\tilde{\mathbf{x}} = (\tilde{x}_1 \ldots \tilde{x}_N)^{\mathrm{T}}$, allowing for a linear scaling in terms of the number of qubits and features. The elements of $\tilde{\mathbf{x}}$ can then be loaded as angles of the rotation gates in the feature map with unitary operator $\hat{U}_{\tilde{\mathbf{x}}}$:

$$\hat{U}_{\tilde{\mathbf{x}}} = \prod_{i=1}^{L_{\mathbf{x}}} \hat{U}_{\mathrm{ent}} \left[\bigotimes_{j=1}^{N} R_{\hat{P}}\left(\tilde{x}_j\right) \right] \tag{1}$$

where $\hat{P} \in \{\hat{X}, \hat{Y}, \hat{Z}\}$ are the Pauli operators, $L_{\mathbf{x}}$ is the number of iterative layers of the data-encoding circuit template, and $\hat{U}_{\mathrm{ent}}$ is an optional entangling layer between rotational gates. We will henceforth refer to the process of mapping $\tilde{\mathbf{x}}$ to the data-encoding circuit via angle encoding as *fingerprint encoding* (figure 1(a)). Angle encoding has been explored as a flexible means of constructing feature maps in data-encoding circuits for machine learning tasks with real-world datasets on near-term devices, for example, in the ZZFeatureMap scheme [55]. However, as explored in greater detail in section 5, we argue that angle encoding of highly compressed data is generally a poor strategy in QML workflows, largely due to the lack of transferability associated with the mapping of unbounded compressed features into bounded angular amplitudes of rotational quantum gates, thus necessitating alternative encoding schemes for molecular QML tasks that are also qubit- and gate-efficient.

3. QMSE

3.1. Hybrid Coulomb-adjacency matrix

The Coulomb matrix is commonly used as an intuitive molecular descriptor [56] that encodes the electrostatic interaction between pairs of atoms (α, β) within molecules. The diagonal represents a fit of atomic energies to nuclear charge data, while the off-diagonal elements scale with the interatomic distance $r_{\alpha\beta}$ as $1/r_{\alpha\beta}$. In this work, we use the *hybrid Coulomb-adjacency* matrix, where we replace $r_{\alpha\beta}$ with the dimensionless parameter $b_{\alpha\beta} \in \{1, 2, 3\}$ depending on the order of the covalent bond. Thus, the modified hybrid Coulomb-adjacency matrix $M_{\alpha\beta}$ is:

$$M_{\alpha\beta} = \begin{cases} 0.5\,\epsilon_T\,\mathcal{Z}_\alpha^d, & \alpha = \beta \\ \dfrac{\epsilon_D \mathcal{Z}_\alpha \mathcal{Z}_\beta}{b_{\alpha\beta}}, & \alpha \neq \beta,\ (\alpha, \beta) \in \mathcal{B} \\ 0, & \alpha \neq \beta,\ (\alpha, \beta) \notin \mathcal{B}, \end{cases} \tag{2}$$

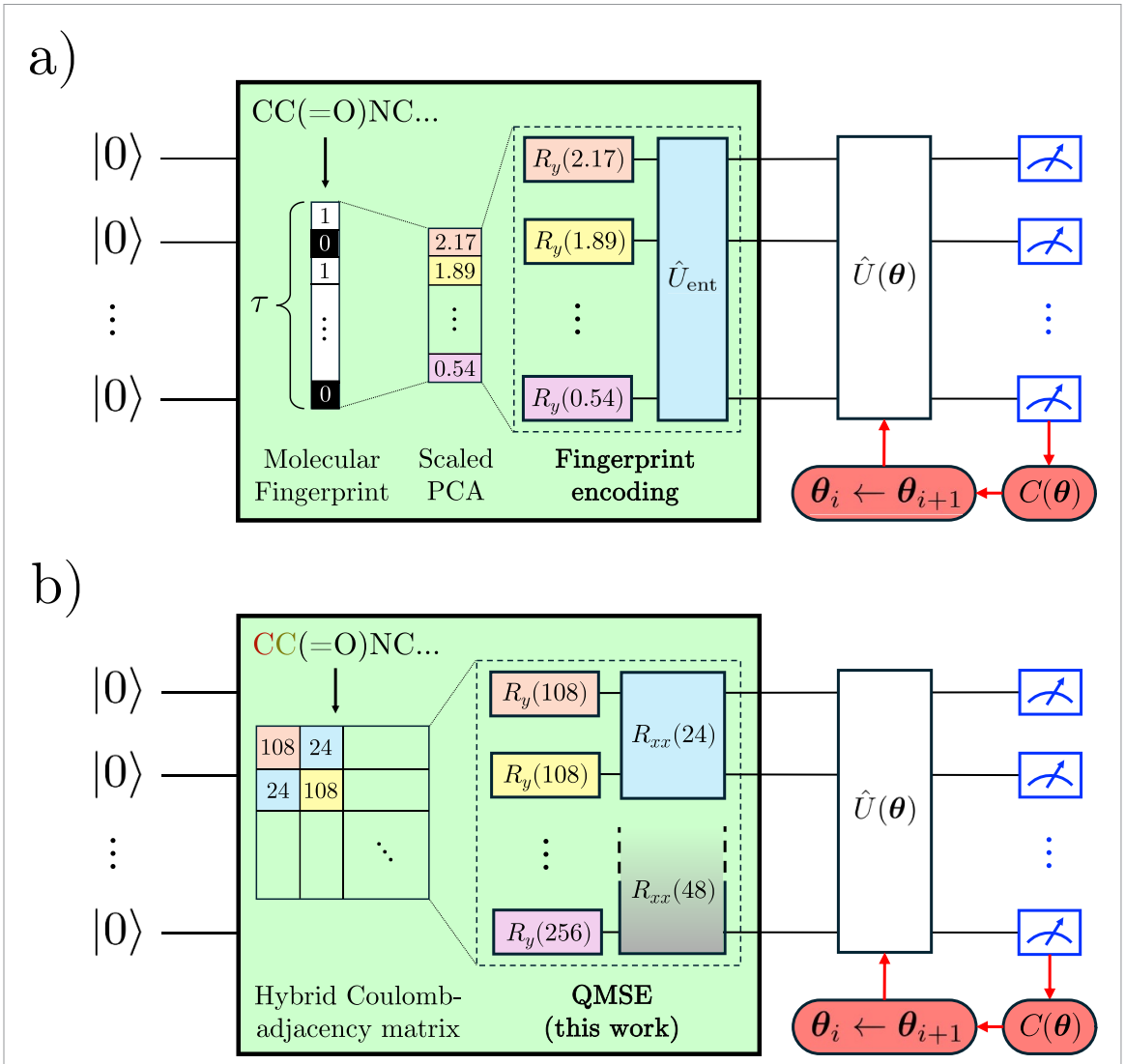

Figure 1. Schematic of the variational QML workflow for two encoding strategies. (a) Fingerprint (angle) encoding: compressed molecular fingerprints are loaded onto the data-encoding layer (in green) as angular rotations. After subsequent evolution of a parametrised ansatz with unitary operator $U(\boldsymbol{\theta})$, the circuit's expectation values are evaluated by measuring an observable $\hat{H}$, and the resulting cost function $C(\boldsymbol{\theta})$ is fed back to a classical optimiser to update the parameters of the ansatz until either the maximum number of iterations or the convergence criterion has been reached. (b) Quantum molecular structure encoding: SMILES strings are instead first converted into a hybrid Coulomb-adjacency matrix and encoded into the quantum circuit by a dedicated data-encoding layer; the ansatz, measurement of $\hat{H}$, and classical parameter-update loop are then applied.

where $\mathcal{Z}$ is the atomic number and $b_{\alpha\beta}$ is the bond order defined in the bond set $\mathcal{B}$. The optional parameters ϵ_D and ϵ_T can be specified to differentiate geometric and optical isomers respectively: for the former, $\epsilon_D = 1$ if a given double bond adopts an E configuration and $\epsilon_D = -1$ if it adopts a Z configuration, while for the latter $\epsilon_T = 1$ if a given tetrahedral atom is assigned an R configuration and $\epsilon_T = -1$ if it is assigned an S configuration.

The hybrid Coulomb-adjacency matrix differs from the canonical Coulomb matrix in three main aspects, namely:

- The off-diagonal elements are non-zero only if atoms α and β possess a covalent bond between them in the molecular bond set $\mathcal{B}$, as opposed to the canonical Coulomb encoding, where the off-diagonal matrix elements are generally non-zero due to the long-range interactions of the Coulomb potential, regardless of whether α and β are covalently bonded. Therefore, this choice reduces the demands of qubit connectivity when preparing the respective data-encoding quantum circuits.
- Using $b_{\alpha\beta}$ in place of $r_{\alpha\beta}$ gives rise to off-diagonal elements with larger magnitudes, enabling the data-encoding rotation gates to be much more sensitive and thus facilitating a greater separation between quantum states. Using $b_{\alpha\beta}$ is also much simpler from an implementation standpoint, whereas $r_{\alpha\beta}$ requires an evaluation of equilibrium bond lengths from standard electronic structure methods to sufficient accuracy.

- The exponent of the diagonal elements, d, is empirically set to 3.0 instead of the commonly used value of 2.4. In our tests, this change was shown to increase the separation of the encoded wave function of molecules (described in further detail in section 3.3).

Thus, we propose hybrid Coulomb-adjacency matrices as a more efficient way to prepare quantum states in QML pipelines.

3.2. Quantum circuit representation

In QMSE, molecules within a given chemical dataset are geometrically represented in the data-encoding quantum circuit layer for QML tasks, by mapping the hybrid Coulomb-adjacency matrix as a sequence of one- and two-qubit gates, similar to a QAOA ansatz circuit for connected graphs (figure 1(b)). As QMSE depends heavily on the atomic identities and covalent connectivity of each molecule, it is expected to produce more distinct representations of the molecular structure compared to fingerprint encoding. The unitary operator that describes the QMSE quantum circuit for a given molecule is:

$$\hat{U}_{\mathbf{x}} = \prod_{k=1}^{L_{\mathbf{x}}} \left\{ \bigotimes_{N_\alpha < N_\beta}^{|\mathcal{B}|} R_{\hat{P}^2}\left(M_{\alpha\beta}\right) \bigotimes_{i=1}^{N_\alpha} R_{\hat{P}}\left(M_{ii}\right) \right\} \tag{3}$$

where $R_{\hat{P}} \in \{R_x, R_y, R_z\}$ and $R_{\hat{P}^2} \in \{R_{xx}, R_{yy}, R_{zz}\}$ are the one- and two-qubit parametrised rotation gates representing the atoms and bonds of the molecule, respectively. It should be noted that N should be greater than or equal to the length of the encoded SMILES string. We now prove the following:

Theorem. *There exists a bijection $\phi : \mathcal{S} \to \mathcal{Q}$ between the set of SMILES strings $\mathcal{S}$ and the set of QMSE unitaries $\mathcal{Q}$ for a given $L_{\mathbf{x}}$ and N.*

Proof. It is easy to see that ϕ is surjective, since for every constructed QMSE unitary operator there exists at least one corresponding SMILES string. For injectivity, suppose $\phi(s_1) = \phi(s_2)$ where $s_1, s_2 \in \mathcal{S}$, or equivalently $\hat{U}_{s_1} = \hat{U}_{s_2}$, where $\hat{U}_{s_1}, \hat{U}_{s_2} \in \mathcal{Q}$. This result implies that the two QMSE unitaries representing s_1 and s_2 are composed of the same $R_{\hat{P}}$ and $R_{\hat{P}^2}$ with the same angular values. As each encoded rotational angle $\tilde{x}$ is not periodic, i.e. $\tilde{x}_1 \neq \tilde{x}_2 - a\pi$ for some $a \in \mathbb{Z}$, all encoded angles are uniquely interspersed between $[-2\pi, 2\pi)$. Hence, each one- and two-qubit rotation gate uniquely represent each atom and bond of a given molecule respectively. Moreover, since the corresponding $R_{\hat{P}^2}$ operators commute with one another, the ordering of the two-qubit gates does not matter even if a different permutation is chosen to populate the two-qubit gates. Therefore, $\hat{U}_{s_1}$ and $\hat{U}_{s_2}$ must necessarily encode the same SMILES string, i.e. $s_1 = s_2$. Thus, ϕ is bijective. □

Furthermore, the choice of $\hat{P}^2$ in the construction of the two-qubit quantum gates is also useful as it allows for rearrangement during transpilation to produce the same wavefunction with a lower circuit depth. This effect can be illustrated in figure 2 with a four-qubit QMSE circuit for the molecule (*E*)-but-2-ene with a SMILES string representation of CC/ = /CC. From equation 2, the hybrid Coulomb-adjacency matrix with a qubit ordering of $|q_0 q_1 q_2 q_3\rangle$ of (*E*)-but-2-ene can be written as:

$$M_{\alpha\beta} = \begin{bmatrix} 108 & 36 & 0 & 0 \\ 36 & 108 & 18 & 0 \\ 0 & 18 & 108 & 36 \\ 0 & 0 & 36 & 108 \end{bmatrix}. \tag{4}$$

Designating R_y and R_{xx} as the encoded one- and two-qubit quantum gates respectively, the encoded R_{xx} gate between qubits q_2 and q_3 can be transpiled such that it can be run simultaneously with the R_{xx} gate between q_0 and q_1 without changing $\hat{U}_{\mathbf{x}}$, due to the commutativity of the R_{xx} gates.

3.3. Properties of QMSE quantum circuits

From the previous theorem, as a bijection exists between the unitary operators of QMSE and SMILES strings, QMSE unitaries have similar properties with those of SMILES strings, and by extension, classical Coulomb and hybrid-Coulomb adjacency matrices. Notably, QMSE operators are invariant to molecular translations and rotations in $SO(3)$; however, they are not permutationally invariant, as reordering the atom-mapped qubits results in different unitary matrices [57].

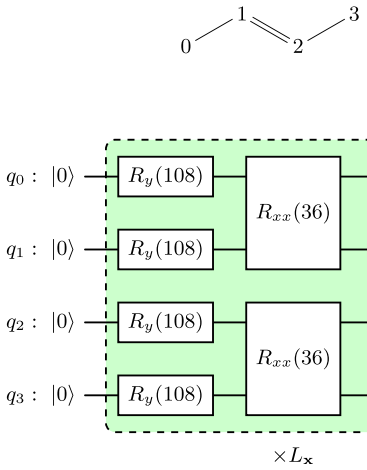

Figure 2. Example molecular encoding layer (bottom) forming part of a 4-qubit quantum circuit representing an (*E*)-but-2-ene molecule (top) using R_y and R_{xx} gates with $L_{\mathbf{x}}$ number of data-encoding layers. The rotational angles associated with the gates are tuned based on the elements of the hybrid Coulomb-adjacency matrix in equation (2).

Another similarity between Coulomb matrices and QMSE operators is the treatment of unrepresented atoms in smaller molecules. Coulomb matrix entries are filled with zeros up to the required maximum number of atoms, while for QMSE, a virtual identity gate is instead implemented for each unrepresented qubit. However, a major difference lies in computational complexity. The matrix entries associated with unrepresented atoms of Coulomb matrices are typically used as input elements in the classical ML regime, while the computational cost of implementing unrepresented qubits in QMSE circuits is essentially zero, as the qubits have already been set to the fiduciary state. This difference produces a highly efficient linear scaling of the combined number of atoms and bonds in a given molecule with the number of data-encoding quantum gates, allowing less complex molecules in a given dataset to be represented by simpler QMSE operators, and vice versa.

A basic problem in cheminformatics is the comparison of chemical moieties between two molecules, which we will label as P and Q for illustration. The comparison can be carried out classically by calculating the chemical similarity. One of the most popular methods involves computing the Tanimoto similarity [58] between the corresponding molecular fingerprints of P and Q, $\mathbf{x}_P$ and $\mathbf{x}_Q$, respectively:

$$T(P,Q) = \frac{|\mathbf{x}_P \cup \mathbf{x}_Q|}{|\mathbf{x}_P| + |\mathbf{x}_Q| - |\mathbf{x}_P \cap \mathbf{x}_Q|}. \tag{5}$$

In the quantum picture, we instead quantify the chemical similarity between P and Q via the quantum overlap or fidelity F between the corresponding wavefunctions $|\psi_P\rangle$ and $|\psi_Q\rangle$:

$$\begin{aligned} F(P,Q) &= |\langle\psi_Q|\psi_P\rangle|^2 \\ &= |\langle\mathbf{0}|\hat{U}_Q^\dagger \hat{U}_P|\mathbf{0}\rangle|^2. \end{aligned} \tag{6}$$

We show that the following property holds for the fidelity of extended molecular chains:

Property. Let the SMILES string representations of P and Q be ordered as $p_1\alpha - \beta p_2$ and $q_1\alpha - \beta q_2$, with common atoms α and β bonded with the same bond order and mapped to the same qubits. Now consider the extended molecular representations of P and Q, $\tilde{P}$ and $\tilde{Q}$ respectively, with some common molecular fragment $\mathbf{c}$ mapped to the same qubit arrangement, i.e. $\tilde{P} = p_1\alpha - \mathbf{c} - \beta p_2$ and $\tilde{Q} = q_1\alpha - \mathbf{c} - \beta q_2$. Then for $L_{\mathbf{x}} = 1$, $F(\tilde{P},\tilde{Q}) = F(P,Q)$.

Proof. Let $\hat{V}$ and $\hat{W}$ be the unitary QMSE representations of the one- and two-qubit rotation layers, respectively. Define:

$$\begin{aligned} \hat{V}_P &= \hat{V}_{p_1} \otimes \hat{V}_\alpha \otimes \hat{V}_\beta \otimes \hat{V}_{p_2} \\ \hat{V}_Q &= \hat{V}_{q_1} \otimes \hat{V}_\alpha \otimes \hat{V}_\beta \otimes \hat{V}_{q_2} \\ \hat{V}_{\tilde{P}} &= \hat{V}_{p_1} \otimes \hat{V}_\alpha \otimes \mathbb{1}_{\mathbf{c}} \otimes \hat{V}_\beta \otimes \hat{V}_{p_2} \\ \hat{V}_{\tilde{Q}} &= \hat{V}_{q_1} \otimes \hat{V}_\alpha \otimes \mathbb{1}_{\mathbf{c}} \otimes \hat{V}_\beta \otimes \hat{V}_{q_2} \\ \hat{U}_c &= \hat{W}_{\mathbf{c}\setminus\{\alpha,\beta\}} \hat{V}_c \end{aligned}$$

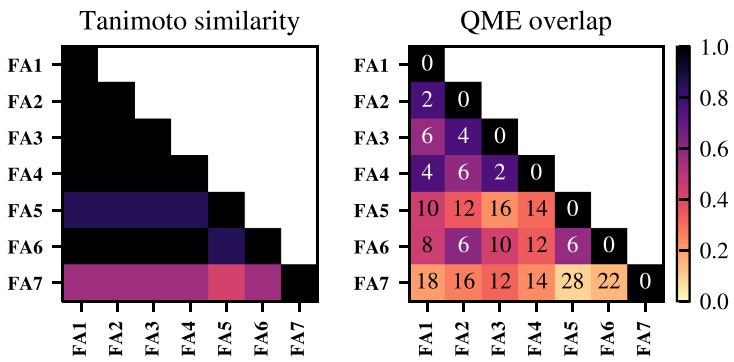

Figure 3. Tanimoto similarity (left) and QMSE fidelity (right) of the fatty acids series **FA1—FA7**. The chemical overlap is computed from SMILES strings encoded via QMSE, using the default $R_y + R_{xx}$ combination. For each overlap pair, the number of qubits of the unitary circuits is also reported after chain contraction.

It follows that

$$\begin{aligned} F\left(\tilde{P},\tilde{Q}\right) &= |\langle\tilde{\mathbf{0}}|\hat{U}_{\tilde{Q}}^{\dagger}\hat{U}_{\tilde{P}}|\tilde{\mathbf{0}}\rangle|^2 \\ &= |\langle\tilde{\mathbf{0}}|\hat{V}_{\tilde{Q}}^{\dagger}\hat{W}_{\tilde{Q}\setminus\mathbf{c}}^{\dagger}\hat{U}_{\mathbf{c}}^{\dagger}\hat{W}_{\alpha\mathbf{c}}^{\dagger}\hat{W}_{\beta\mathbf{c}}^{\dagger}\hat{W}_{\beta\mathbf{c}}\hat{W}_{\alpha\mathbf{c}}\hat{U}_{\mathbf{c}}\hat{W}_{\tilde{P}\setminus\mathbf{c}}\hat{V}_{\tilde{P}}|\tilde{\mathbf{0}}\rangle|^2 \\ &= |\langle\tilde{\mathbf{0}}|\hat{V}_{\tilde{Q}}^{\dagger}\hat{W}_{\tilde{Q}}^{\dagger}\hat{W}_{\tilde{P}}\hat{V}_{\tilde{P}}|\tilde{\mathbf{0}}\rangle|^2 \\ &= |\langle\mathbf{0}|\hat{V}_{Q}^{\dagger}\hat{W}_{Q}^{\dagger}\hat{W}_{P}\hat{V}_{P}|\mathbf{0}\rangle|^2 \\ &= F(P,Q)\,. \end{aligned}$$

□

This property is instrumental in setting up the following corollary for computing the fidelity between molecules with common substructures using a reduced number of qubits:

Corollary. *If two given molecular representations $\tilde{P}$ and $\tilde{Q}$ share some common molecular fragments $\mathcal{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots)$ bonded identically to the same common atoms mapped to the same qubits, then $F(\tilde{P},\tilde{Q})$ with N qubits can be evaluated with $N - N_{\mathcal{C}}$ qubits via $F(P,Q)$ by eliminating $\mathcal{C}$ in both $\tilde{P}$ and $\tilde{Q}$.*

We will refer to the process set out in the corollary as *chain contraction*. This procedure is guaranteed for QMSE circuits with $L_x = 1$, regardless of the choice of $R_{\hat{p}}$ and $R_{\hat{p}_2}$. To highlight the effectiveness of chain contraction, we numerically computed the quantum fidelities of a curated series of seven unsaturated fatty acids labelled **FA1—FA7**, each containing 34 carbon atoms (figure 3), and compared the result to the classical Tanimoto similarity (figure 3). Although the Tanimoto metric was generally able to distinguish **FA7**, with a larger degree of unsaturation, from the rest of the structures, the long saturated chains of **FA1—FA6** make them difficult to distinguish. Quantum fidelities are more effective in evaluating chemical similarity, reflecting the wide variability of unsaturation and the double bond positions. Although the evaluation of quantum fidelities would ordinarily be computationally intensive due to the large number of atoms per molecule, eliminating common molecular fragments via chain contraction allows for a potentially large reduction in the number of implemented qubits and quantum gates. Figure 3 illustrates the number of qubits required to calculate each fidelity pair. The structures of **FA1—FA7** and the chain contraction procedure are outlined in appendix A.

4. Datasets

We compiled a dataset of 105 linear saturated small organic molecules from the CRC Handbook of Chemistry and Physics (95th Edition) [59]. This dataset includes 50 alkanes, 38 monohydric alcohols, and 17 monohydric ethers with varying degrees of positional isomerism. Canonical SMILES for each molecule were first constructed via the RDKit canonicalisation algorithm. The alcohol and ether SMILES strings were then reordered with the oxygen atom in the left-most position, to maximise fidelities between similar chemical moieties and vice versa. To assess the performance of the algorithm with datasets of increasing size and complexity, we further partitioned the complete dataset of 105 molecules into two subsets: the *alkane* subdataset with only the alkane structures, and the *oxygen* subdataset with only the alcohol and ether structures.

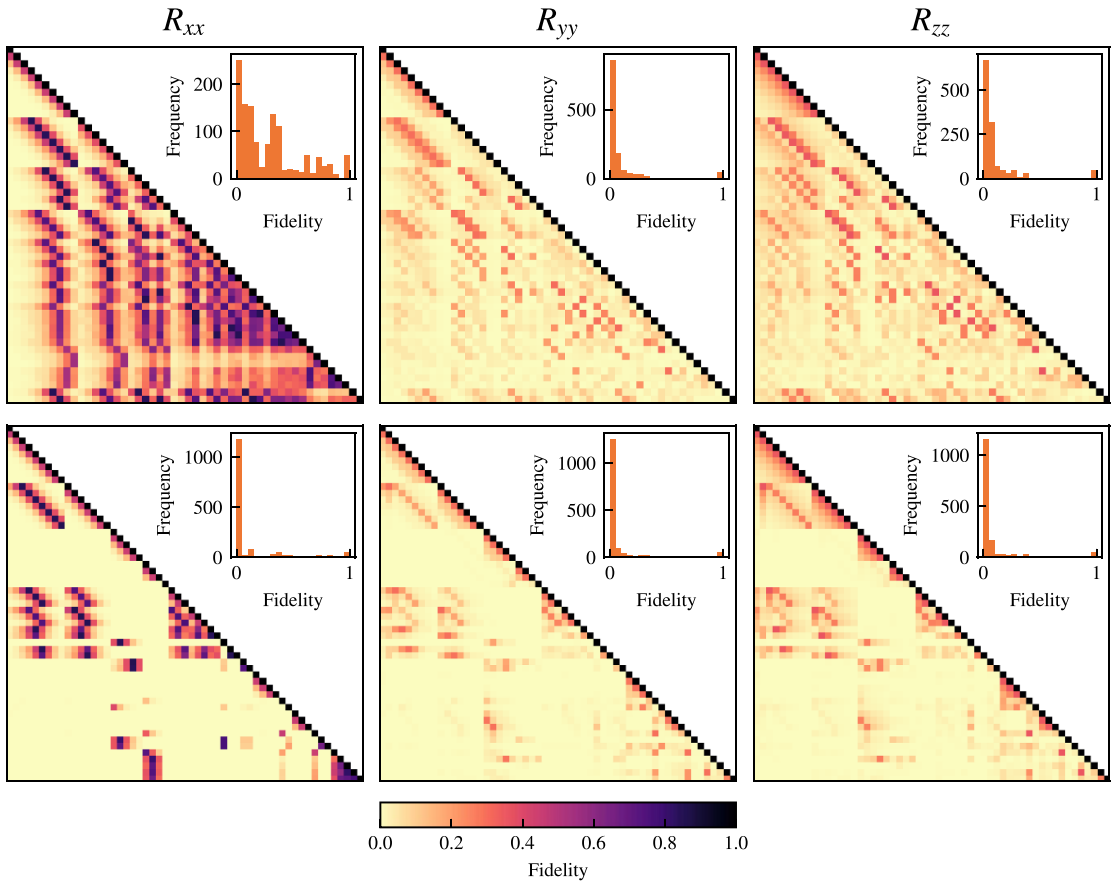

Figure 4. Heatmaps of fidelity matrices for chain-contracted molecules encoded with QMSE within the alkane- (top row) and oxygen- (bottom row) subdatasets. The one-qubit data-encoding gate for all configurations is fixed as R_y, while the two-qubit data-encoding gate is varied to produce the fidelity matrix for R_{xx} (left column), R_{yy} (middle column), and R_{zz} (right column). The colour corresponds to the fidelity of each molecule pair within their respective insets. Based on the overall distribution of the fidelity values, $R_y + R_{xx}$ is selected as the default setup for QMSE.

To better understand chemical similarity within the QMSE framework, we systematically benchmarked the effect of different combinations of one- and two-qubit rotation gates on the fidelities of the chemical datasets, as defined in equation 5. We fixed the single-qubit rotation to R_y gates and varied the two-qubit entangling operations among R_{xx}, R_{yy}, and R_{zz}. All simulations were performed with $L_{\mathbf{x}} = 1$ under noiseless statevector conditions for different pairs of chain-contracted molecules within the alkane- and oxygen-subdatasets.

Figure 4 shows the heatmaps of quantum fidelity matrices for the alkane- and oxygen-subdatasets. In both cases, the $R_y + R_{xx}$ configuration displays the widest range of fidelities, compared to$R_y + R_{yy}$ or $R_y + R_{zz}$. This separation directly translates into improved discrimination capability in QML tasks. Accordingly, we use R_y for single-qubit rotations and R_{xx} for two-qubit entangling gates as our default encoding setup for all subsequent experiments using QMSE.

5. Results

We perform three main QML numerical experiments to investigate the effectiveness of QMSE, summarised in table 1. The first task involves a binary classification of the alkane subdataset, where we predict whether a given molecule is in the gas phase at 100 °C. We contrast the default $R_y + R_{xx}$ QMSE configuration (Runs 3–4) with standard fingerprint encoding (Runs 1–2). For the latter runs, the data was preprocessed by converting the subdataset SMILES strings into RDKit topological molecular fingerprints comprising 2048 bits each, and subsequently reducing each fingerprint via PCA into ten coordinates, corresponding to the same maximum qubit size of the data-encoding layer of QMSE. The fingerprint encoding layer was then prepared with a single layer of R_y gates and rotation angles corresponding to the PCA-reduced coordinates scaled in the range $[-2\pi, 2\pi]$, followed by a linear entangling layer of CNOT gates. Both the fingerprint-encoded and QMSE data-encoding layers are followed by a variational ansatz composed of R_y gates followed by an entangling layer $\hat{U}_{\text{ent}}$, where $\hat{U}_{\text{ent}}$ is either a linear arrangement of

Table 1. Summary of classification and regression runs with the variational approach of section 5. For the data-encoding layer, the configuration used for fingerprint encoding employs an initial R_y rotation gate layer followed by a linear chain of CNOT gates with $L_{\mathbf{x}} = 1$; for QMSE the configuration utilises the default $R_y + R_{xx}$ configuration with $L_{\mathbf{x}} = 1$. From left to right, the columns indicate the run ID, type of task assigned to the machine learning model (classification or regression), molecular dataset, ansatz configuration, Hamiltonian, and the maximum number of COBYLA iterations.

				Ansatz configuration					Iterations
Run ID	Task	Dataset	Encoding	1-qubit	2-qubit	Entanglement	Layers	Hamiltonian	($/10^3$)
1	Classification	Alkane	Fingerprint	R_y	CZ	Linear	$[1-5]$	Global all-$\hat{Z}$	1
2	Classification	Alkane	Fingerprint	R_y	CZ	Pairwise	$[1-5]$	Global all-$\hat{Z}$	1
3	Classification	Alkane	QMSE	R_y	CZ	Linear	$[1-5]$	Global all-$\hat{Z}$	1
4	Classification	Alkane	QMSE	R_y	CZ	Pairwise	$[1-5]$	Global all-$\hat{Z}$	1
5	Classification	Complete	QMSE	R_y	CZ	Pairwise	$[1-5]$	Global all-$\hat{Z}$	2
6	Classification	Complete	QMSE	R_y	CRX	Pairwise	$[1-5]$	Global all-$\hat{Z}$	2
7	Classification	Complete	QMSE	R_y	CRX	Pairwise	$[1-5]$	*IIIIZZIIII*	2
8	Classification	Complete	QMSE	R_y	CRX	Pairwise	$[1-5]$	*IIIZZZZIII*	2
9	Classification	Complete	QMSE	R_y	CRX	Pairwise	$[1-5]$	*ZZIIIIIIII*	2
10	Regression	Alkane	QMSE	R_y	CRX	Pairwise	$[1-6]$	Global all-$\hat{Z}$	10
11	Regression	Alkane	QMSE	R_y	CRX	Full	$[1-6]$	Global all-$\hat{Z}$	10

CZ gates (Runs 1, 3) or a pairwise arrangement of CZ gates (Runs 2, 4):

$$\hat{U}(\boldsymbol{\theta}) = \prod_{i=1}^{L_{\boldsymbol{\theta}}} \hat{U}_{\text{ent}} \left[\bigotimes_{j=1}^{N} R_y\left(\theta_j\right) \right], \tag{7}$$

where $\hat{U}_{\text{ent}}$ can be configured to express an ansatz with full, linear, and pairwise entanglement, and L_θ is the number of ansatz layers.

The variational quantum classifier (VQC) circuit is then measured in the global all-$\hat{Z}$ basis and optimised via a gradient-free COBYLA regime for a maximum of 1000 iterations with an L_2 loss function, assigning $y = 1$ for predicted $\langle \hat{H} \rangle$ values greater than 0 and $y = 0$ for predicted $\langle \hat{H} \rangle$ values less than 0.

To further evaluate the performance of QMSE for a wider range of chemical moieties, we broaden the same classification task to the complete dataset, including monohydric alcohol and ether molecules (Runs 5–9), and increase the maximum number of COBYLA iterations to 2000 for improved convergence. As the classification task complexity is increased, we compare the effect of the variational ansatz entangling CZ gate (Run 5) with a more expressive controlled-R_x (CRX) gate (Run 6) with pairwise entangling configurations. We also vary the Hamiltonian measured by the VQC circuit by experimenting with local Hamiltonians, where we measure a selection of qubits in the $\hat{Z}$-basis (Run 7–9), as local cost functions have been shown to be more trainable than global cost functions in parametrised quantum circuits up to $L_\theta \in \mathcal{O}(\log N)$ [60]. In particular, we wish to explore the light-cone phenomenon arising from the pairwise entangling arrangement of the ansatz [61], whose effect is more pronounced when measuring the middle two qubits (Run 7) and the middle four qubits (Run 8) in the default $\hat{Z}$-basis. As the oxygen subdataset contains alcohols and ethers that predominantly feature the oxygen atom in the first two qubit positions, we also seek to understand the impact of measuring the first two qubits in the $\hat{Z}$-basis (Run 9) compared to the other local Hamiltonian runs.

Finally, we tackle the much more difficult task of regression using the alkane subdataset to predict boiling points via a variational quantum regressor (VQR). Here, we normalise the boiling points (in Kelvin) to the range $[-0.5, 0.5]$, rather than $[-1, 1]$, as the difficulty for the QML model to express expectation values with larger magnitudes is much higher. Moderating the range of predicted expectation values reduces the risk of underfitting molecules with the lowest or highest boiling points. This choice also allows the model to potentially extrapolate boiling points outside the predicted range. We benchmark the performance with an R_y + CRX variational ansatz, with either a pairwise (Run 10) or a full entanglement (Run 11) configuration. The VQR circuit was then measured in the global all-$\hat{Z}$ basis and optimised for a maximum number of 10 000 COBYLA iterations with an L_2 loss function.

All runs were evaluated via stratified k-fold cross validation, with the alkane and complete datasets split into $k = 5$ equally sized groups of samples. The classification tasks were assessed with the mean accuracy scores of both training and test datasets, while the regression tasks were evaluated with the coefficient of determination R^2 of both training and test datasets. For each cross-validated iteration,

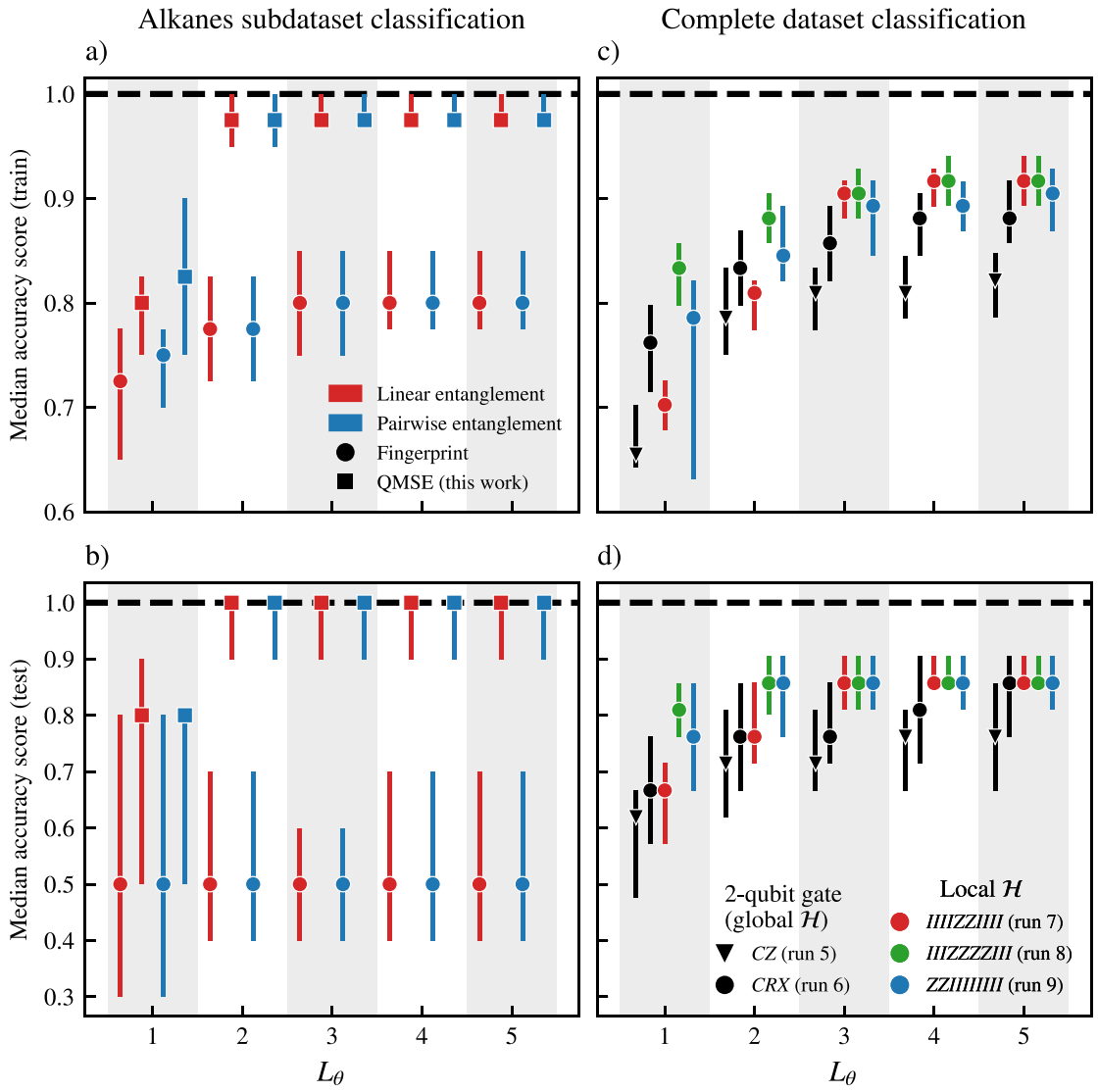

Figure 5. Accuracy scores when classifying molecules in the alkane (left, (a) and (b) and complete (right, (c) and (d) datasets for Runs 1−9 with 1−5 L_θ number of ansatz layers. The top row shows the median training accuracy scores, and the bottom row shows the median test accuracy scores. The error bars indicate the 16$^{\text{th}}$ and 84$^{\text{th}}$ percentile values of the average accuracies of the k-fold splits.

a total of 100 random initial coordinates $\boldsymbol{\theta}_i \in [-2\pi, 2\pi]$ for the variational ansatz were sampled, and the median results were tabulated across the number of initial coordinates and the number of k-fold samples. This process was repeated across a discrete range of ansatz layers for each run, with the classification tasks ranging between 1–5 ansatz layers and the regression task between 1 and 6 ansatz layers.

5.1. Classification

For the classification task on the alkane subdataset, QMSE achieves excellent results with a consistently high training and corresponding test accuracy score for increasing L_θ (figures 5(a) and (b)). At $L_\theta = 3$ and above, aside from the data outliers of 2,2,3,3-tetramethylpentane and occasionally methane, the VQC model coupled with QMSE was trained with perfect accuracy scores, and generalised to unseen data in the test splits with minimal overfitting. In contrast, the modest improvement in training accuracy scores for the VQC model coupled fingerprint encoding with increasing L_θ translates poorly to the test splits with low accuracy scores. This result strongly indicates that fingerprint encoding coupled to PCA-reduced data is a poor data encoding framework for representing molecular structures in chemical QML tasks. The stark contrast in performance between fingerprint encoding and QMSE can also be attributed to the difference in their respective loss curves, where QMSE converges to a much lower loss (refer to appendix B for more details on the loss curves). For different ansatz entanglement schemes with both fingerprint encoding and QMSE, there appears to be no significant difference between linear and pairwise entanglement in terms of accuracy scores.

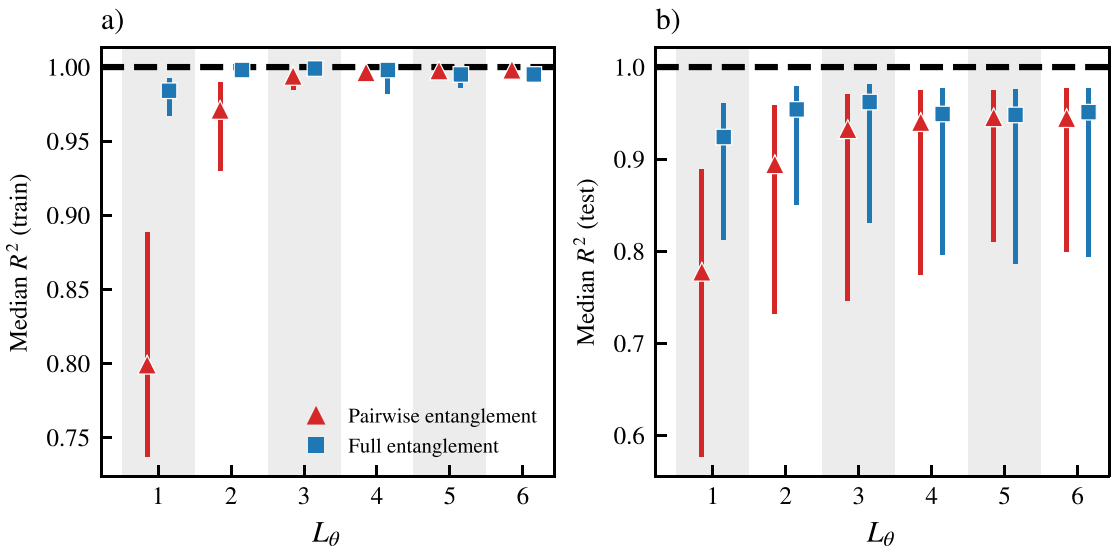

Figure 6. R^2 scores for boiling point prediction of the alkane subdataset for Runs 10−11 with 1−6 L_θ number of ansatz layers, with (a) depicting the median training scores (left), and (b) the median test scores (right). The error bars depict the 16th and 84th percentile values of the average R^2 values of the k-fold splits.

Extending our exploration to the classification of the complete dataset, we observe good performance in the training of the VQC model with QMSE and generalisation to the test samples (figures 5(c) and (d)). Owing to the added complexity of the expanded dataset with alcohol and ether moieties, we found that modifying the ansatz entangling gate from CZ (Run 5) to CRX (Run 6) produced a marked increase in training and test accuracy results. Modification of the Hamiltonian from a global all $\hat{Z}$-basis to a local $\hat{Z}$-basis further improved performance at higher L_θ. This result is particularly impressive, especially from a quantum error perspective, as reducing the number of measured qubits also reduces the source of noise attributed to crosstalk between qubits [62]. In particular, Runs 7–8 attain slightly higher training accuracies of over 90% at $L_\theta = 5$ compared to Run 9, suggesting that the light-cone phenomenon is a little more significant in training VQCs than the emphasis in measuring the first two qubits.

5.2. Regression

Lastly, by considering the alkane regression task (figure 6), we find excellent training R^2 scores, however the generalisation of the trained VQR model to the test data reaches a limit with increasing L_θ. This result suggests a proliferation of local minima with lower R^2 scores arising from different starting ansatz coordinates having a larger effect on the more difficult task of predicting continuous variables. Nevertheless, this is still an encouraging result with R^2 test values exceeding 0.95 for optimal starting conditions with little to no overfitting. In terms of ansatz arrangement, Run 11 with full entanglement appears to perform slightly better in terms of both R^2 training and test scores compared to Run 10 with pairwise entanglement, which is unsurprising given the increase in parameters for the same L_θ.

6. Discussion

Our results show that QMSE greatly outperforms conventional fingerprint encoding in the alkane classification task, in addition to performing quite well for more complex classification and regression tasks. The improved gains observed with QMSE arise, not only from improved state separability, but also from its inherent interpretable structure, mirroring key advantages that are also desirable in classical machine learning. In classical ML, interpretability enables understanding of how input features affect model decisions, guiding both model debugging and feature engineering. For QML, where feature maps are implemented as quantum circuits, the data-encoding scheme governs the entire hypothesis space, and thus understanding its structure is especially critical in designing more expressive and highly trainable models.

The interpretability of QMSE can be rationalised by considering two main decision boundaries maximising separation between datapoints with different properties (figure 7). Consider a molecular dataset encoded with QMSE, with their resulting statevectors residing in the Hilbert space $\mathcal{H}_N$. Now consider the Hilbert space $\mathcal{H}_a = \mathcal{H}_{N-1} \otimes \mathcal{H}_{|0\rangle}$, such that $\mathcal{H}_a \subset \mathcal{H}_N$. The molecular statevectors

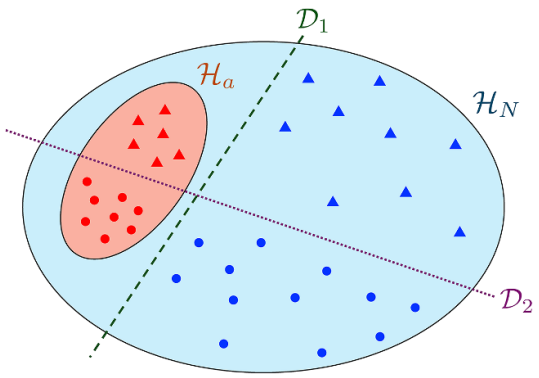

Figure 7. Illustration of decision boundaries $\mathcal{D}_1$ and $\mathcal{D}_2$ separating a molecular datatset encoded with QMSE in $\mathcal{H}_N$. $\mathcal{D}_1$ (green line) aims to separate the datapoints encoding smaller molecules in $\mathcal{H}_a$ (red) from the larger molecules outside $\mathcal{H}_a$ (blue), while $\mathcal{D}_2$ (purple line) seeks to separate datapoints based on atomic or bond identities (circles vs. triangles).

in $\mathcal{H}_a$ with at least one less encoded atom than the maximum number of encoded qubits can thus be separated from the maximally encoded statevectors lying outside $\mathcal{H}_a$ with the decision boundary $\mathcal{D}_1$, and this reasoning extends for increasingly small subspaces of $\mathcal{H}_a$ that smaller molecules reside in. To further distinguish between atomic and bond identities, a second decision boundary $\mathcal{D}_2$ can be used that cuts across $\mathcal{H}_a$ and $\mathcal{H}_N \cup \mathcal{H}_a'$. For example, consider some molecular dataset $S = \{|\psi_{\text{CC}}\rangle, |\psi_{\text{CCC}}\rangle, |\psi_{\text{C=C}}\rangle, |\psi_{\text{OC}}\rangle, |\psi_{\text{C=CC}}\rangle, |\psi_{\text{OCC}}\rangle\}$ in $\mathcal{H}_N$ for $N = 3$. $\mathcal{D}_1$ can be established between molecular subsets with two atoms and three atoms, i.e. $S_1 = \{|\psi_{\text{CC}}\rangle, |\psi_{\text{C=C}}\rangle, |\psi_{\text{OC}}\rangle\}$ and $S_2 = \{|\psi_{\text{CCC}}\rangle, |\psi_{\text{C=CC}}\rangle, |\psi_{\text{OCC}}\rangle\}$, where $S_1 \subset \mathcal{H}_a$ and $S_2 \subset \mathcal{H}_N \cup \mathcal{H}_a'$. For $\mathcal{D}_2$, the datapoints can be further partitioned based on the bond order of the first C−C bond, as well as the identity of the first atom.

Quantum state overlaps of different molecules derived from QMSE exhibit higher variance and wider spread compared to those from fingerprint encodings (figures 3 and 4), reflecting better state distinguishability and fewer kernel concentration issues [37]. This result suggests that QMSE defines a more meaningful quantum feature space, consistent with theoretical prescriptions for quantum advantage [44].

Finally, the reduced two-qubit depth and support for fidelity-preserving chain contraction ensure that the circuits remain executable on noisy or early fault-tolerant hardware. Thus, the physical basis of QMSE greatly facilitates a good balance between interpretability, expressivity, and noise robustness.

7. Conclusion

In summary, we have developed QMSE as a highly effective data-encoding strategy for representing molecular structures in QML classification and regression tasks. Feature encoding techniques, such as fingerprint encoding, suffer from poor generalisability arising from the mapping of compressed data as rotational amplitudes. In contrast, QMSE provides a straightforward means of segregating datapoints via the construction of decision boundaries, which maximises state separation in terms of the corresponding chemical moieties.

We propose several future avenues for expanding the use of QMSE in practical quantum computing applications. The first theme focuses on broadening QMSE to encode other data structures beyond organic molecules in drug discovery, such as periodic unit cells of crystalline materials. Graph embeddings of crystal structures have demonstrated considerable success in classical machine learning workflows for accelerating materials discovery [63–65]. Hence, QMSE could enable the prediction of crystalline material properties via QML. Due to the innate ability of QMSE to load classical data linearly in the form of SMILES strings as data-encoding quantum circuits, variations of QMSE can be conceptualised that optimise the loading of other types of string information, such as encoding text as embedded tokens in quantum natural language processing [66]. Furthermore, the synthesis of QMSE circuits from the linear composition of one- and two-qubit quantum gates can be exploited in generative artificial intelligence (genAI) frameworks [67], to produce quantum circuits that can be mapped procedurally back into molecular structures.

We also seek to optimise the various QML algorithms that are compatible with QMSE. Expanding on this work for variational QML models, such as VQC and VQR, we aim to improve on ansätze by considering the evaluation of the Shapley values of their parameters, thus enhancing interpretability and synergistic effects with QMSE circuits [68]. To combat the traditional problems associated with variational

quantum algorithms, such as vanishing gradients and an abundance of local minima with poor solutions, non-variational quantum algorithms, such as quantum kernels, QSVM and quantum graph neural network models, can be used instead [38, 69–71]. We expect this approach to be especially powerful when combined with chain contraction, allowing for the efficient evaluation of fidelities from different pairs of encoded molecular wavefunctions, and providing opportunities to perform QML tasks on more complex chemical data inventories.

We would also like to briefly outline our future approach in addressing the exponential concentration problem of encoding increasingly complex datapoints for higher N. While we have shown that QMSE greatly improves meaningful correlation between data-encoding unitaries and chemical moieties relative to other standard quantum encoding schemes such as basis, amplitude, and angle encoding, we acknowledge that a more in-depth study of overlap kernel matrices between datapoint pairs needs to be carried out to specifically address the phenomenon of exponential concentration. In a similar vein to the previous point of studying other QML models, we seek to combine QMSE circuits with data reuploading techniques for higher $L_{\mathbf{x}}$ that further aim to curtail exponential concentration, thus potentially enhancing their generalisation capabilities significantly [54, 72].

Furthermore, while SMILES strings can be uniquely mapped into their respective QMSE operators, the same molecule may be represented using different SMILES strings, and hence with different QMSE circuits. Thus, it is imperative during data preprocessing to order the atoms of the SMILES strings of the molecular dataset consistently, such that the optimal comparison between datapoints is maximised. Other potential schemes to address this shortcoming could be the randomisation of choosing a SMILES string representing each molecule in the dataset to reduce selection bias, the inclusion of more datapoints of different SMILES strings mapped to the same molecule for better training alignment, or the conversion of more sophisticated more string representations such as the International Chemical Identifier [73] into QMSE data-encoding circuits. We aim to examine this problem more systematically with curated databases in the future, such as subsets of GDB-17 that feature interesting organic scaffolds for potential drug discovery applications [74].

Finally, we consider the position of QMSE as an effective data-encoding method in the context of the ongoing transition from near-term to early FTQC regimes. In the early fault-tolerant regime, error-corrected logical qubits enable high-fidelity preparation of molecular graph-state encodings via block-encoding of adjacency or Coulomb-adjacency matrices, directly mapping connectivity into entanglement patterns [75]. Reduced noise and parallelisable CZ-based graph-state circuits allow deeper variational ansätze without barren plateaux, improving gradient magnitudes and convergence [76, 77]. The usage of partial quantum error correction methods to prepare high-quality, arbitrary R_z rotations gates that are outside of the Clifford gate set via magic state injection also allow larger variational quantum circuits to be implemented [78]. Another example involves the parallelisation of evaluating separate expectation values of individual datapoints [79, 80], thus further tightening integration between quantum devices and high-performing computers in addition to leveraging the latter for QEC methods. The preparation of fault-tolerant graph states also reduces depth overhead by commutativity, enhancing the resilience to residual errors [81, 82]. Consequently, QML models based on explicit graph-state encodings are expected to exhibit faster convergence and better generalisation on early FTQC hardware compared to their pre-FTQC counterparts [83]. Overall, QMSE is expected to benefit significantly from early FTQC frameworks, regardless of combination with variational or non-variational QML models.

Data availability statement

We provide a publicly accessible GitHub repository (https://github.com/stfc/quantum-molecular-encodings) hosting the routines for mapping SMILES strings to hybrid Coulomb-adjacency matrices, as well as generation of their corresponding data-encoding quantum circuits. The repository also contains examples formatted as Jupyter notebooks. We also provide the 105-molecule dataset from the 95$^{\text{th}}$ CRC Handbook of Chemistry and Physics [59] with canonical SMILES and normalised bond-order matrices. The implementation of the molecular structure encoding layer introduced in this work will be made available in the main Qiskit Machine Learning library [84] from version 0.9. The code and data produced by this work are distributed under the (CC BY) license without any warranty.

Acknowledgments

EA and CB are grateful to M Emre Sahin for helping to draft the code that inputs quantum circuits as feature maps in Qiskit Machine Learning. We acknowledge helpful conversations with Jason Crain and help from Edward O. Pyzer-Knapp and Benjamin C B Symons during the initial set-up of the project.

This work was supported by the Hartree National Centre for Digital Innovation, a UK Government-funded collaboration between STFC and IBM. IBM, the IBM logo, and www.ibm.com are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. The current list of IBM trademarks is available at www.ibm.com/legal/copytrade.

Author contributions

We provide the author contributions following the CRediT (Contributor Roles Taxonomy) scheme. CB and DJW conceptualised the molecular encoding scheme. CB and EA contributed equally to developing the QML methodology, software, formal analysis, investigation and writing of the original draft of the manuscript. EA produced the figures and integrated the QMSE framework with Qiskit Machine Learning. DM and CB compiled the datasets and performed the data curation. IT oversaw project administration and supervision. SM was responsible for funding acquisition and project administration. All the authors have reviewed and provided feedback on the final manuscript.

Appendix A. Simulating quantum fidelities of fatty acids

The procedure for simulating the quantum fidelities of the unsaturated fatty acid series **FA1 — FA7** is outlined in this section. The molecules and structures are described in table 2 and figure 8, respectively. The SMILES string representations of the fatty acid series were first canonicalised via RDKit and subsequently reordered with the carboxylic acid moieties left-aligned, so as to ensure maximum structural overlap between the molecules. Using the chain contraction procedure, the maximum overlap between pairs of SMILES strings was omitted and the quantum fidelity circuit was constructed for the molecule pair via QMSE with R_y and R_{xx} as the rotational and entangling gates, respectively, and $L_{\mathbf{x}} = 1$. The number of qubits required to construct the circuit was determined by the length of the longer reduced SMILES string. As the fatty acids display geometric isomerism from the C=C double bonds in the *Z* configuration, the optional argument ϵ_T was imposed on the required R_y one-qubit rotations. The resulting fidelity values are shown in figure 3.

Table 2. Summary of the identities and SMILES representation orderings of the fatty acid series *FA*1 — *FA*7.

ID	PubChem CID	IUPAC Name	SMILES Representation
FA1	52 921 804	19*Z*,22*Z*,25*Z*,28*Z*-tetratriacontate traenoic acid	OC(=O)CCCCCCC CCCCCCCCCC/C=C\C /C=C\C/C=C\C/C=CCCCCC
FA2	14 753 668	19*Z*,22*Z*,25*Z*,28*Z*,31*Z*-tetratriacont apentaenoic acid	OC(=O)CCCCCCCCC CCCCCCC/C=C\C/C=C\C /C=C\C/C=C\C/C=C\CC
FA3	52 921 817	16*Z*,19*Z*,22*Z*,25*Z*,28*Z*,31*Z*-tetratriacontahexaenoic acid	OC(=O)CCCCCCCCC CCCCC/C=C\C/C=C\C/C=C\C /C=C\C/C=C\C/C=C\CC
FA4	52 921 824	16*Z*,19*Z*,22*Z*,25*Z*,28*Z*-tetratriacontap entaenoic acid	OC(=O)CCCCCCCC CCCCCC/C=C\C/C=C\C /C=C\C/C=C\C/C=CCCCCC
FA5	92 033 288	25*Z*-tetratriaco ntenoic acid	OC(=O)CCCCCCCCCCCCC CCCCCCCCC/C=C\CCCC CCCC
FA6	171 118 569	25*Z*,28*Z*,31*Z*-tetratriacon tatrienoic acid	OC(=O)CCCCCCCCCCCC CCCCCCCCCC/C=C\C /C=C\C/C=C\CC
FA7	171 117 702	4*Z*,7*Z*,10*Z*,13*Z*,16*Z*,19*Z*,22*Z*,25*Z*,28*Z*,31*Z*-tetratriacon tadecenoic acid	OC(=O)CC/C=C\C/C=C\C/ C=C\C/C=C\C/C=C\C/C=C\C /C=C\C/C=C\C/C=C\C/C=C\CC

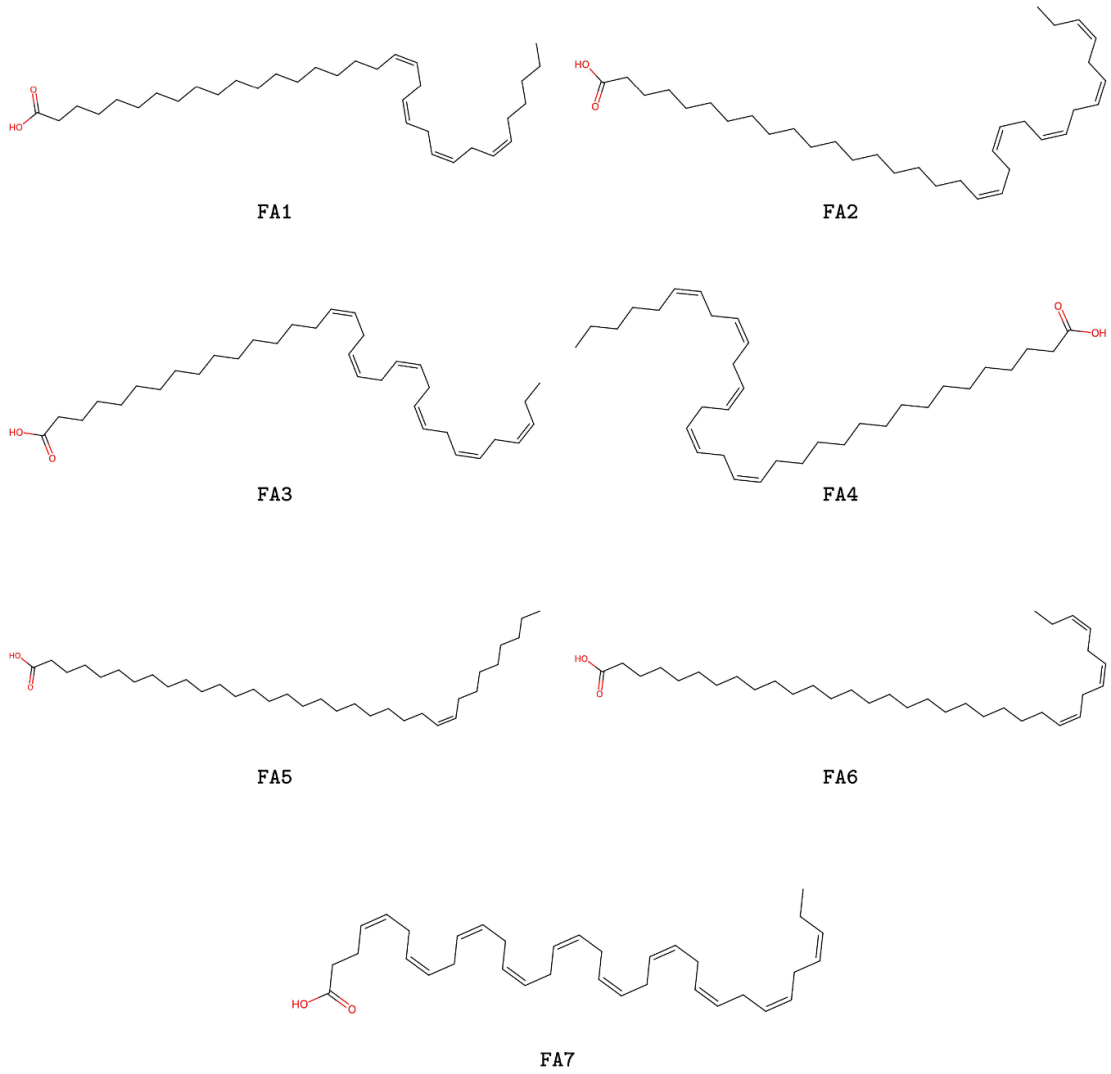

Figure 8. Skeletal structures of the fatty acid series **FA1 – FA7** as outlined in table 2.

Appendix B. Loss curves of alkane subdataset VQC model

Figure 9 shows the training L_2 loss curves for the VQC model of the alkane subdataset tasks for Runs 1–4 and different k-fold splits. In terms of the ansatz, random parameter initialisation with similar losses were selected in this illustration for all runs with $L_\theta = 5$. We found that all five k-fold splits behave similarly, and QMSE (Runs 3–4) exhibits superior convergence in model performance compared to fingerprint encoding (Runs 1–2), consistent with the improvement in training accuracies as shown in figure 5.

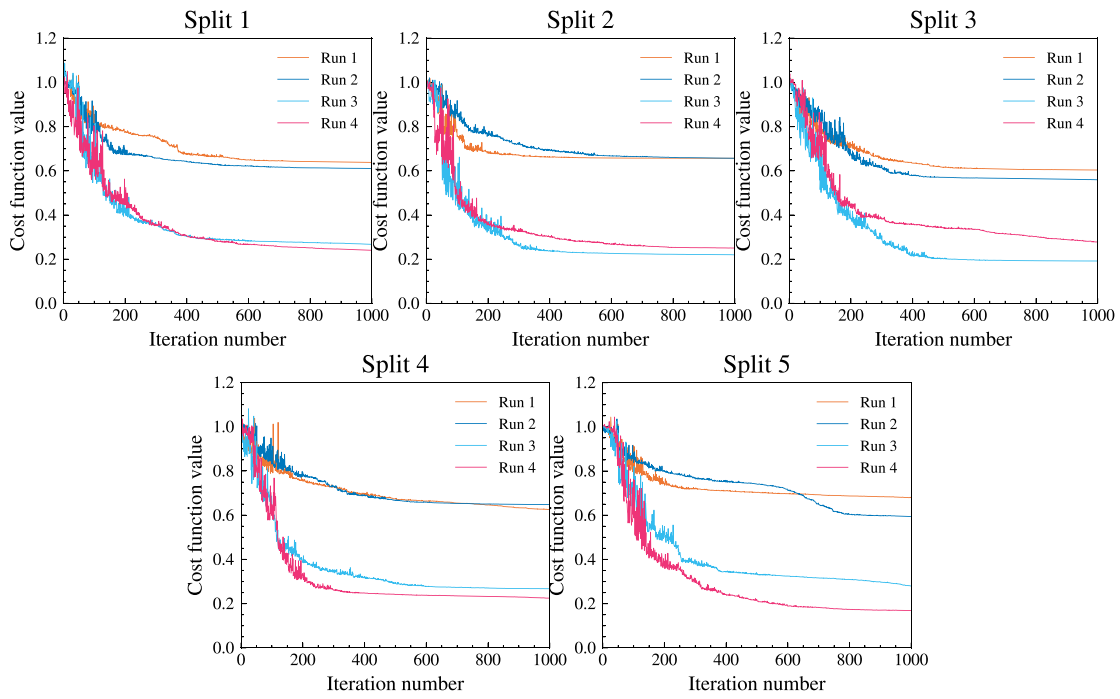

Figure 9. Training L_2 loss curves on the VQC model for Runs 1–4 on the alkane subdataset and their respective k-fold splits, with random ansatz parameter initialisation and $L_\theta = 5$.

Appendix C. Forecasts for near-term devices

As shown in section 4, the strength of QMSE as an encoding scheme for QML stems from its ability to discriminate molecules of similar type. To evaluate the utility of executing QMSE on currently available quantum hardware, we performed fidelity measures for the alkane- and oxygen-subdatasets using noise profiles designed to mimic the errors present in quantum processors. The quantum circuits constructed via equation (3) were transpiled to `ibm_pittsburgh` device, an IBM quantum chip built using superconducting transmon qubits arranged in a heavy-hexagonal topology. A representative noise model for this device was selected based on calibration data recorded on the 10th of August 2025 at 18:59:40 UTC. This model was incorporated to sampling of the fidelity measurements via IBM's Qiskit Aer simulator. Each fidelity measure was sampled over 10 000 shots. Heatmaps of quantum fidelity matrices for these noisy simulations are shown in figure 10.

The presence of noise is expected to lead to a degradation of the main diagonal elements in the heatmap matrix compared to the noiseless heatmaps of figure 4. However, we note that the difference in the fidelities between the noiseless and the noisy cases is minute, which can be mainly attributed to the shallow entangling gate depths of the circuits produced by the QMSE scheme enhanced with chain contraction.

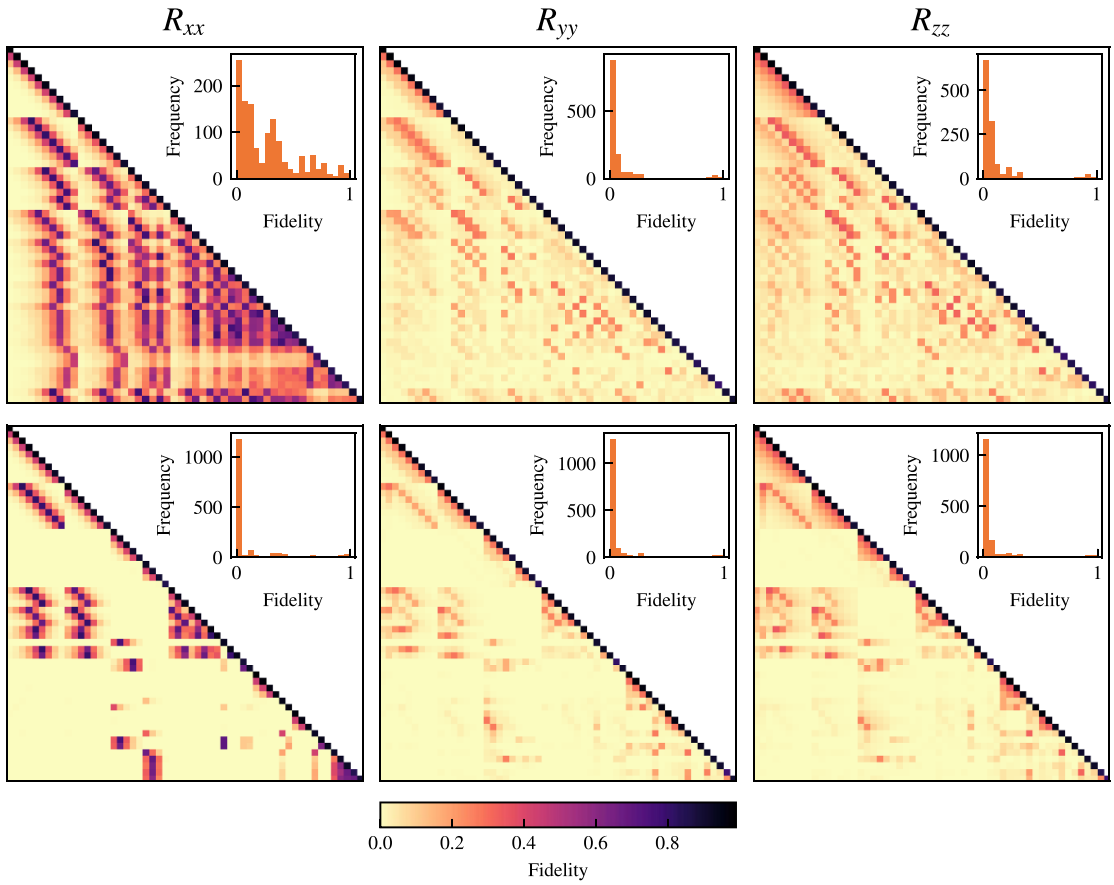

Figure 10. Heatmaps of fidelity matrices for chain-contracted molecules encoded with QMSE within the alkane- (top row) and oxygen- (bottom row) subdatasets, calculated in the presence of device noise. The noise profile corresponds to data for the `ibm_pittsburgh` device.

ORCID iDs

Choy Boy 0000-0002-6782-9433
Edoardo Altamura 0000-0001-6973-1897
Dilhan Manawadu 0000-0002-3575-8060
Ivano Tavernelli 0000-0001-5690-1981
Stefano Mensa 0000-0002-0938-144X
David J Wales 0000-0002-3555-6645

References

[1] Jumper J *et al* 2021 Highly accurate protein structure prediction with alphafold *Nature* **596** 583–9
[2] Huang E T, Yang J-S, Liao K Y, Tseng W C, Lee C, Gill M, Compas C, See S and Tsai F-J 2024 Predicting blood–brain barrier permeability of molecules with a large language model and machine learning *Sci. Rep.* **14** 15844
[3] Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N and Lloyd S 2017 Quantum machine learning *Nature* **549** 195–202
[4] Mensa S, Sahin E, Tacchino F, Barkoutsos P K and Tavernelli I 2023 Quantum machine learning framework for virtual screening in drug discovery: a prospective quantum advantage *Mach. Learn.: Sci. Technol.* **4** 015023
[5] Li W *et al* 2024 A hybrid quantum computing pipeline for real world drug discovery *Sci. Rep.* **14** 16942
[6] Sundaram A 2025 Challenges and opportunities in quantum computing in healthcare *Quantum Computing for Healthcare Data* (Academic) pp 91–118
[7] Somoroff A, Ficheux Q, Mencia R A, Xiong H, Kuzmin R and Manucharyan V E 2023 Millisecond coherence in a superconducting qubit *Phys. Rev. Lett.* **130** 267001
[8] Tuokkola M, Sunada Y, Kivijärvi H, Albanese J, Grönberg L, Kaikkonen J, Vesterinen V, Govenius J and Möttönen M 2025 Methods to achieve near-millisecond energy relaxation and dephasing times for a superconducting transmon qubit *Nat. Commun.* **16** 5421
[9] Réglade U *et al* 2024 Quantum control of a cat qubit with bit-flip times exceeding ten seconds *Nature* **629** 778–83
[10] Miao K C *et al* 2023 Overcoming leakage in quantum error correction *Nat. Phys.* **19** 1780–6
[11] Tripathi V, Chen H, Khezri M, Yip K-W, Levenson-Falk E and Lidar D A 2022 Suppression of crosstalk in superconducting qubits using dynamical decoupling *Phys. Rev. Appl.* **18** 024068
[12] Cong I, Choi S and Lukin M D 2019 Quantum convolutional neural networks *Nat. Phys.* **15** 1273–8

[13] Yamasaki H, Subramanian S, Sonoda S and Koashi M 2020 Learning with optimized random features: exponential speedup by quantum machine learning without sparsity and low-rank assumptions *Adv. Neural Inf. Process. Syst.* vol **33** pp 13674–87
[14] Hibat-Allah M, Mauri M, Carrasquilla J and Perdomo-Ortiz A 2024 A framework for demonstrating practical quantum advantage: comparing quantum against classical generative models *Commun. Phys.* **7** 68
[15] Riofrio C A, Mitevski O, Jones C, Krellner F, Vuckovic A, Doetsch J, Klepsch J, Ehmer T and Luckow A 2024 A characterization of quantum generative models *ACM Trans. Quantum Comput.* **5** 1–34
[16] Gil-Fuster E, Eisert J and Bravo-Prieto C 2024 Understanding quantum machine learning also requires rethinking generalization *Nat. Commun.* **15** 2277
[17] Gupta H, Varshney H, Sharma T K, Pachauri N and Verma O P 2022 Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction *Complex Intell. Syst.* **8** 3073–87
[18] Schuld M and Killoran N 2022 Is quantum advantage the right goal for quantum machine learning? *PRX Quantum* **3** 030101
[19] Boulougouri M, Vandergheynst P and Probst D 2024 Molecular set representation learning *Nat. Mach. Intell.* **6** 754–63
[20] Potdar K, Pardawala T S and Pai C D 2017 A comparative study of categorical variable encoding techniques for neural network classifiers *Int. J. Comput. Appl.* **175** 7–9
[21] Nuñez-Andrade E, Vidal-Daza I, Ryan J W, Gómez-Bombarelli R and Martin-Martinez F J 2025 Embedded machine-readable molecular representation for resource-efficient deep learning applications *Digit. Discovery* **4** 776–89
[22] Cao P-Y, He Y, Cui M-Y, Zhang X-M, Zhang Q and Zhang H-Y 2024 Group graph: a molecular graph representation with enhanced performance, efficiency and interpretability *J. Cheminf.* **16** 133
[23] Smaldone A M *et al* 2025 Quantum machine learning in drug discovery: applications in academia and pharmaceutical industries *Chem. Rev.* **125** 5436–60
[24] Möttönen M, Vartiainen J J, Bergholm V and Salomaa M M 2005 Transformation of quantum states using uniformly controlled rotations *Quantum Inf. Comput.* **5** 467–73
[25] Araujo I F, Park D K, Petruccione F and da Silva A J 2021 A divide-and-conquer algorithm for quantum state preparation *Sci. Rep.* **11** 6329
[26] Shende V, Bullock S and Markov I 2006 Synthesis of quantum-logic circuits *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **25** 1000–10
[27] Xia R and Kais S 2020 Hybrid quantum-classical neural network for calculating ground state energies of molecules *Entropy* **22** 828
[28] Kiss O, Tacchino F, Vallecorsa S and Tavernelli I 2022 Quantum neural networks force fields generation *Mach. Learn.: Sci. Technol.* **3** 035004
[29] Le I N M, Kiss O, Schuhmacher J, Tavernelli I and Tacchino F 2025 Symmetry-invariant quantum machine learning force fields *New J. Phys.* **27** 023015
[30] Boiko D A, Reschützegger T, Sanchez-Lengeling B, Blau S M and Gomes G 2025 Advancing molecular machine learning representations with stereoelectronics-infused molecular graphs *Nat. Mach. Intell.* **7** 771–81
[31] Weinhold F and Landis C R 2001 Natural bond orbitals and extensions of localized bonding concepts *Chem. Educ. Res. Pract.* **2** 91–104
[32] Rupp M, Tkatchenko A, Müller K-R and Lilienfeld O A V 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
[33] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev. B* **87** 184115
[34] Heid E, Greenman K P, Chung Y, Li S-C, Graff D E, Vermeire F H, Wu H, Green W H and McGill C J 2023 Chemprop: a machine learning package for chemical property prediction *J. Chem. Inf. Model.* **64** 9–17
[35] Torabian E and Krems R V 2025 Molecular representations of quantum circuits for quantum machine learning (arXiv:2503.05955)
[36] Larocca M, Thanasilp S, Wang S, Sharma K, Biamonte J, Coles P J, Cincio L, McClean J R, Holmes Z and Cerezo M 2025 Barren plateaus in variational quantum computing *Nat. Rev. Phys.* **7** 174–89
[37] Thanasilp S, Wang S, Cerezo M and Holmes Z 2024 Exponential concentration in quantum kernel methods *Nat. Commun.* **15** 5200
[38] Glick J R, Gujarati T P, Corcoles A D, Kim Y, Kandala A, Gambetta J M and Temme K 2024 Covariant quantum kernels for data with group structure *Nat. Phys.* **20** 479–83
[39] Kamata Y, Tran Q H, Endo Y and Oshima H 2025 Molecular quantum transformer (arXiv:2503.21686)
[40] Peruzzo A, McClean J, Shadbolt P, Yung M-H, Zhou X-Q, Love P J, Aspuru-Guzik A and O'brien J L 2014 A variational eigenvalue solver on a photonic quantum processor *Nat. Commun.* **5** 4213
[41] Cervera-Lierta A, Kottmann J S and Aspuru-Guzik A 2021 Meta-variational quantum eigensolver: learning energy profiles of parameterized hamiltonians for quantum simulation *PRX Quantum* **2** 020329
[42] Li G, Zhao X and Wang X 2024 Quantum self-attention neural networks for text classification *Sci. China Inf. Sci.* **67** 142501
[43] Tran L D, Nguyen S M and Arai M 2020 GAN-based noise model for denoising real images *Computer Vision - Accv 2020: 15th Asian Conf. on Computer Vision, (Kyoto, Japan, 30 November–4 December 2020), (Revised Selected Papers, Part IV)* (Springer) pp 560–72
[44] Liu Y, Arunachalam S and Temme K 2021 A rigorous and robust quantum speed-up in supervised machine learning *Nat. Phys.* **17** 1013–7
[45] Thanasilp S, Wang S, Nghiem N A, Coles P and Cerezo M 2023 Subtleties in the trainability of quantum machine learning models *Quantum Mach. Intell.* **5** 21
[46] Crognaletti G, Grossi M and Bassi A 2025 Estimates of loss function concentration in noisy parametrized quantum circuits (arXiv:2410.01893)
[47] Bermejo P, Braccia P, Rudolph M S, Holmes Z, Cincio L and Cerezo M 2024 Quantum convolutional neural networks are (effectively) classically simulable (arXiv:2408.12739)
[48] Tang E 2021 Quantum principal component analysis only achieves an exponential speedup because of its state preparation assumptions *Phys. Rev. Lett.* **127** 060503
[49] Huang H-Y, Kueng R and Preskill J 2021 Information-theoretic bounds on quantum advantage in machine learning *Phys. Rev. Lett.* **126** 190505
[50] Brown T *et al* 2020 Language models are few-shot learners *Adv. Neural Inf. Process. Syst.* vol **33** pp 1877–901
[51] Li Q, Huang Y, Hou X, Li Y, Wang X and Bayat A 2024 Ensemble-learning error mitigation for variational quantum shallow-circuit classifiers *Phys. Rev. Res.* **6** 013027
[52] Nilakantan R, Bauman N, Dixon J S and Venkataraghavan R 1987 Topological torsion: a new molecular descriptor for SAR applications comparison with other descriptors *J. Chem. Inf. Comput. Sci.* **27** 82–85

[53] Balewski J, Amankwah M G, Beeumen R V, Bethel E W, Perciano T and Camps D 2024 Quantum-parallel vectorized data encodings and computations on trapped-ion and transmon qpus *Sci. Rep.* **14** 3435
[54] Pérez-Salinas A, Cervera-Lierta A, Gil-Fuster E and Latorre J I 2020 Data re-uploading for a universal quantum classifier *Quantum* **4** 226
[55] Havlíček V, Córcoles A D, Temme K, Harrow A W, Kandala A, Chow J M and Gambetta J M 2019 Supervised learning with quantum-enhanced feature spaces *Nature* **567** 209–12
[56] Neese F 2003 An improvement of the resolution of the identity approximation for the formation of the coulomb matrix *J. Comput. Chem.* **24** 1740–7
[57] Montavon G, Hansen K, Fazli S, Rupp M, Biegler F, Ziehe A, Tkatchenko A, Lilienfeld A and Müller K-R 2012 Learning invariant representations of molecules for atomization energy prediction *Advances in Neural Information Processing Systems* vol 25 (Curran Associates, Inc.) pp 440–8
[58] Bajusz D, Rácz A and Héberger K 2015 Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **7** **20**
[59] Haynes W M ed 2014 *CRC Handbook of Chemistry and Physics* 95th edn (CRC Press)
[60] Cerezo M, Sone A, Volkoff T, Cincio L and Coles P J 2021 Cost function dependent barren plateaus in shallow parametrized quantum circuits *Nat. Commun.* **12** 1791
[61] Abedi E, Beigi S and Taghavi L 2023 Quantum lazy training *Quantum* **7** 989
[62] Parrado-Rodríguez P, Ryan-Anderson C, Bermudez A and Müller M 2021 Crosstalk suppression for fault-tolerant quantum error correction with trapped ions *Quantum* **5** 487
[63] Cheng G, Gong X-G and Yin W-J 2022 Crystal structure prediction by combining graph network and optimization algorithm *Nat. Commun.* **13** 1492
[64] An R, Xie C, Chu D, Li F, Pan S and Yang Z 2024 A machine-learning-assisted crystalline structure prediction framework to accelerate materials discovery *ACS Appl. Mat. Inter.* **16** 36658–66
[65] Luo X, Wang Z, Gao P, Lü J, Wang Y, Chen C and Ma Y 2024 Deep learning generative model for crystal structure prediction *npj Comput. Mat.* **10** 254
[66] Varmantchaonala C M, Fendji J L K E, Schöning J and Atemkeng M 2024 Quantum natural language processing: a comprehensive survey *IEEE* **12** 99578–98
[67] Lin X, Xia Y, Li Y, Huang Y-P, Liu S, Zhang J and Gao Y Q 2025 In-silico 3D molecular editing through physics-informed and preference-aligned generative foundation models *Nat. Commun.* **16** 6043
[68] Heese R, Gerlach T, Mücke S, Müller S, Jakobs M and Piatkowski N 2025 Explaining quantum circuits with shapley values: Towards explainable quantum machine learning *Quantum Mach. Intell.* **7** 27
[69] Sancho-Lorente T, Román-Roche J and Zueco D 2022 Quantum kernels to learn the phases of quantum matter *Phys. Rev. A* **105** 042432
[70] Gil-Fuster E, Eisert J and Dunjko V 2024 On the expressivity of embedding quantum kernels *Mach. Learn.: Sci. Technol.* **5** 025003
[71] Piperno S, Ceschini A, Chang S Y, Grossi M, Vallecorsa S and Panella M 2025 A study on quantum graph neural networks applied to molecular physics *Phys. Scr.* **100** 065126
[72] Rodriguez-Grasa P, Ban Y and Sanz M 2025 Neural quantum kernels: training quantum kernels with quantum neural networks *Phys. Rev. Res.* **7** 023269
[73] Krenn M *et al* 2022 SELFIES and the future of molecular string representations *Patterns* **3** 100588
[74] Ruddigkeit L, van Deursen R, Blum L C and Reymond J-L 2012 Enumeration of 166 billion organic small molecules in the chemical Universe database GDB-17 *J. Chem. Inf. Model.* **52** 2864–75
[75] Hein M, Eisert J and Briegel H J 2004 Multiparty entanglement in graph states *Phys. Rev. A* **69** 062311
[76] Cerezo M *et al* 2025 Does provable absence of barren plateaus imply classical simulability? *Nat. Commun.* **16** 7907
[77] Preskill J 2018 Quantum computing in the nisq era and beyond *Quantum* **2** 79
[78] Dangwal S, Vittal S, Seifert L M, Chong F T and Ravi G S 2025 Variational quantum algorithms in the era of early fault tolerance (arXiv:2503.20963)
[79] Cattelan M, Yarkoni S and Lechner W 2025 Parallel circuit implementation of variational quantum algorithms *npj Quantum Inf.* **11** 27
[80] Tsukayama D, ichi Shirakashi J, Shibuya T and Imai H 2025 Enhancing computational accuracy with parallel parameter optimization in variational quantum eigensolver *AIP Adv.* **15** 015226
[81] Gottesman D 1997 Stabilizer codes and quantum error correction (arXiv:9705052)
[82] Huang J, Lewis L, Huang H-Y and Preskill J 2024 Predicting adaptively chosen observables in quantum systems (arXiv:2410.15501)
[83] Marshall S C, Gyurik C and Dunjko V 2023 High dimensional quantum machine learning with small quantum computers *Quantum* **7** 1078
[84] Sahin M E, Altamura E, Wallis O, Wood S P, Dekusar A, Millar D A, Imamichi T, Matsuo A and Mensa S 2025 Qiskit machine learning: an open-source library for quantum machine learning tasks at scale on quantum hardware and classical simulators (arXiv:2505.17756)