Quantum Information Processing (2023) 22:435
https://doi.org/10.1007/s11128-023-04188-7

REVIEW PAPER

The barren plateaus of quantum neural networks: review, taxonomy and trends

Han Qi[1] · Lei Wang[1] · Hongsheng Zhu[1] · Abdullah Gani[2] · Changqing Gong[1]

Received: 8 August 2023 / Accepted: 21 November 2023 / Published online: 11 December 2023

Abstract

In the noisy intermediate-scale quantum (NISQ) era, the computing power displayed by quantum computing hardware may be more advantageous than classical computers, but the emergence of the barren plateau (BP) has hindered quantum computing power and cannot solve large-scale problems. This summary analyzes the phenomenon of the BP in the quantum neural network that is rapidly developing in the NISQ era. This article will review the research status of the BP problem in the quantum neural network (QNN) in the past five years from the analysis of the source of the BP, the current stage solution, and the future research direction. First of all, the source of the BP was briefly explained and then classified the BP solution from different perspectives, including quantum embedding in QNN, ansatz parameter selection and structural design, and optimization algorithms. Finally, the BP problem in the QNN is summarized, and the research direction for solving problems in the future is made.

Keywords Quantum neural network · Barren plateau · Quantum computing

✉ Han Qi
qihan@sau.edu.cn

Lei Wang
wang.lei.98@qq.com

Hongsheng Zhu
821199361@qq.com

Abdullah Gani
abdullah@um.edu.my

Changqing Gong
gongchangqing@sau.edu.cn

1 School of Computer Science, Shenyang Aerospace University, Shenyang 110000, People's Republic of China

2 Faculty of Computing and Informatics, University of Malaya, Jalan Universiti, Kuala Lumpur 50603, Malaysia

1 Introduction

Over the last decade, there have been a large number of successes in quantum computing, including quantum simulation, quantum chemistry, and quantum machine learning [1–5]. However, there is still no guarantee that quantum computing will have greater advantages over classical computing in the coming years [6]. In the era of NISQ, VQA has become an important area for quantum computing, and a recent practical algorithm to realize quantum advantages [7]. The VQA is a hybrid quantum-classic algorithm, including a quantum approximation optimization algorithm (QAOA) [8], quantum machine learning (QML) [9], variational quantum eigensolver (VQE) [10], etc., uses parameterized quantum circuits (PQC) to run on a quantum computer, and then optimizes parameters on a classical computer. VQA provides a framework for dealing with various problems. It has been widely used in multiple areas, including finding base states [10–12] and founding states [13–15], mathematical applications [16–18], machine learning [19–21], etc.

As a subset of VQA [22], a small-scale QNN has been implemented in quantum computing for supervised learning tasks [19]. The QNN usually contains three parts, as shown in Fig 1: The classical data are converted into quantum states by the encoding quantum circuit U_x, followed by training PQC $U(\theta)$, and the minimum target is to minimize the parameters in the quantum circuit to minimize the target function f.$U(\theta)$ represents the quantum unitary transform, and θ uses a classical optimizer to update and feedback back to the quantum device, iterated update parameters until they meet the given task. As a quantum promotion of the classical neural network, QNN has benefited from strong quantum computing power to achieve quantum advantages [23–25]and promoted the study of classification models, generating models, and reinforcement learning [26–31]. However, McClean et al. recently found that for QNN with random structure, the gradient would disappear exponentially with the size of the system, resulting in poorer trainability of QNN. This phenomenon of gradient disappearance is called BP [23].

The generation of BP in QNN hinders whether it can achieve quantum advantages. This is because the exponential disappearance of the gradient means that exponential resources are required for optimization, and since the standard goal of quantum algorithms is that with the polynomial scaling of the system (the exponential scaling of classical computers), the generation of BP will cause exponential scaling to destroy the quantum acceleration. As the scale of the problem increases, the quantum circuits become untrainable. Therefore, avoiding or eliminating the impact of BP is an important challenge to prove that QNN is better than classical neural networks in the future.

At present, more and more researchers have noticed the training of QNN and have proposed a variety of strategies to alleviate the Bp. It includes initializing parameter strategy, hierarchical learning strategy, meta-learning strategy, etc., but there is no unified method to solve BP. Because of this, this paper classifies and summarizes various solutions based on the structure of QNN, to facilitate readers to find strategies.

In this review, we will conduct a unified survey of the existing options on BP. In Sect. 2, the source, meaning, and representation method of the BP are allowed to facilitate readers to have a preliminary understanding of the BP. In Sect. 3, we

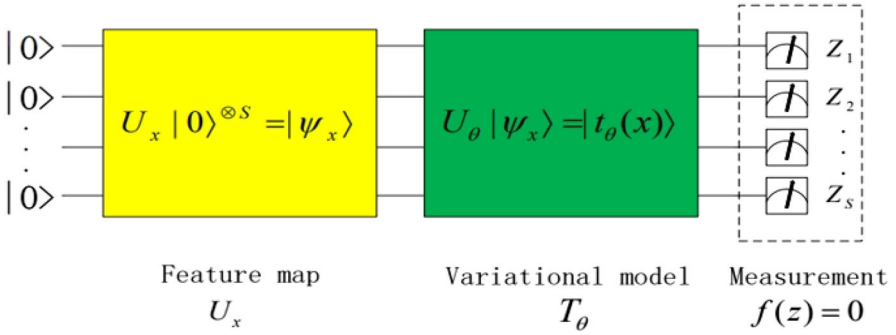

Fig. 1 A schematic illustration of QNN [22]. The classical data encoded by the quantum characteristics mapping method to the Hilbert space to complete the transformation of the classic to a quantum state $|\Psi_x\rangle := U_x|0\rangle^{\otimes s}$. Subsequently, the transformer model ansatz was optimized with PQC, and the quantum state was converted into a new state $|t_\theta(x)\rangle := U_\theta|\Psi_x\rangle$, where the parameters $\theta \in \Theta$ are chosen to minimize a certain loss function. Finally, measurement, classical post-processes of the measurement results $z = (z_0, z_1, ..., z_s)$, and extract the output of the model y:= f(z)

introduce the solution to BP in the past few years. In Sect. 4, we have the development and prospects of the future QNN.

2 Barren plateaus

2.1 The origins of barren plateaus

In the quantum algorithm, many factors lead to BP, including the high expressibility of the quantum circuit [31], a cost function of a global goal [32], as well as excess entanglement in the quantum circuit [33, 34] and noises in the NISQ device [35, 36].

When the cost function exhibits a barren plateau, the cost function gradient vanishes exponentially with the system size. Without loss of generality, we consider here the following generic definition of a barren plateau.

By analyzing the scaling of the variance of the cost function partial derivative, one can detect the presence of barren plateaus in the cost function landscape [30]. We denote this variance as $Var_\theta[\partial_i C]$, where the expectation values in the variance are taken over θ. Specifically, when the cost function exhibits a barren plateau, one finds that $Var_\theta[\partial_i C]$ is exponentially vanishing with the number of qubits, i.e.,

$$Var_\theta[\partial_i C] \leq F(n), with\ F(n) \in o\left(\frac{1}{b^n}\right) \tag{1}$$

for some $b > 1$. Then, combining Eq. (1) with Chebyshev's inequality, the probability that the cost derivative deviates from its mean value (of zero [31]) is bounded as

$$Pr(|\partial_i| \geq c) \leq \frac{Var_\theta[\partial_i C]}{c^2} \leq \frac{F(n)}{c^2} \tag{2}$$

for any $c > 0$. Equation (2) shows that, on average, the cost function partial derivatives will be exponentially small across the landscape, meaning that an exponential precision is needed to estimate the gradients and determine a cost minimizing direction.

McClean, Cerezo et al. explained the phenomenon of the BP, indicating that the BP is related to the cost function. However, little is known about the landscape of the cost function. Holmes et al. used the numerical relationship between the expressibility and trainability of PQC to establish a link between the landscape of the cost function and the assumed expressivity [37]. The stronger the expression of the PQC, the smaller the gradient variance of the cost function, the flatter the landscape, and the more difficult it is to train [38]. Although Holmes et al. have proved the connection between expressiveness and BP, this value proves that it depends on the choice of cost functions and requires more in-depth research on more common situations. The connection between cost function landscapes was further explored by Arrasmith et al. who demonstrated that when a BP phenomenon occurs, the mean value of the exponential cost is concentrated and narrow gorges occur [39] and that the occurrence of two of these phenomena must lead to the other. This means that for a wide range of circuits, instead of expensive gradient calculations to diagnose BP, estimations are made by cost differences.

In 2018, McClean et al. first proposed the BP phenomenon, where gradient disappearance exists during the training of QNN. In their paper, they used a random PQC structure to verify that for a class of PQC satisfying the T-2 design [40, 41], the probability of gradient along any reasonable direction for some fixed accuracy (nonzero) decreases exponentially with the increase in the number of qubits.

$$U(\theta) = U(\theta_1, \ldots \theta_L) = \prod_{l=1}^{L} U_l(\theta_l) W_l \tag{3}$$

where $U_l(\theta_l) = exp(-i\theta_l V_l)$, V_l is a Hermitian operator, and W_l is a generic unitary operator that does not angle θ_l. The objection function $E(\theta)$ is expressed as the exception value over some Hermitian operator H, $E(\theta) = \langle 0| U(\theta) \dagger HU(\theta) |0\rangle$. The gradient-based optimization $E(\theta)$ is represented as:

$$\partial_k E \equiv \frac{\partial E(\theta)}{\partial \theta_k} = i \langle 0| U_-^{\dagger} \left[V_k, U_+^{\dagger} H U_+ \right] U_- |0\rangle \tag{4}$$

where $U_- \equiv \prod_{l=0}^{k-1} U_l(\theta_l) W_l$, $U_+ \equiv \prod_{l=k}^{L} U_l(\theta_l) W_l$. The circuits U_+, U_- are independent distribution, and these quantum circuits are designed to agree with the statistical properties of the Haar measure at the first and second moments, i.e., they realize a second-order approximation to the Haar measure. Using the unitary T-designs, the expectation value of the gradient is 0 when the T-1 design is satisfied, $\partial_k E = 0$. When the T-2 design is satisfied, the gradient variance decays exponentially as the number of qubits increases.

Cerezo et al. have expanded the definition of the BP, and the existence of the BP depends on the cost function [32] and establishes a global cost function (the observation value acts on all qubits) and local cost function (the observation value acts on several qubits) connection between the BP. As shown in Fig. 2, for the use of global cost functions, regardless of the depth of the circuit, the BP phenomenon

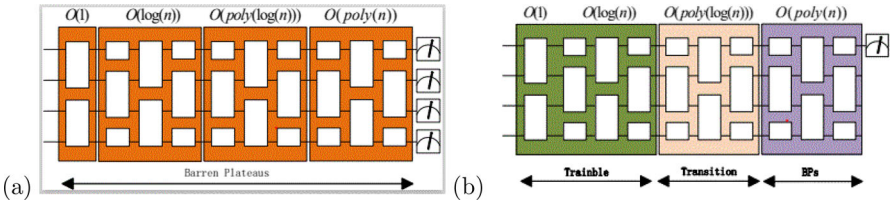

Fig. 2 The trainability of the ansatz in terms of the circuit depth for the global cost function (**a**) and the local cost function (**b**), is modified from [32]

will be present. For the local cost function, when the circuit depth $D \in O\left(log\left(n\right)\right)$, the gradient maintains a maximum polynomial disappearance, which is trained. To facilitate understanding, consider a toy problem corresponding to the target state$|0\rangle$. We assume a tensor product ansatz of the form $V\left(\theta\right) = \otimes_{j=1}^{n} e^{i\theta^j \sigma_x^{(j)}/2}$, with the goal of findings the angles θ_j such that $V\left(\theta\right)|0\rangle = |0\rangle$. Suppose the global cost function is:

$$C_G = Tr\left[O_G V\left(\theta\right)|\phi_0\rangle\langle\phi_0| V\left(\theta\right)^{\dagger}\right] \tag{5}$$

where $O_G = I - |0\rangle\langle 0|$, the global cost function is converted to $C_G = 1 - \prod_{j=1}^{n} \cos^2 \frac{\theta^j}{2}$. The mean value is $\frac{\partial C_G}{\partial \theta^j} = 0$, the gradient concentrates exponentially around zero. The variance of its gradient is $Var\left[\frac{\partial C_G}{\partial \theta^j}\right] = \frac{1}{8}\left(\frac{3}{8}\right)^{n-1}$, which vanishes exponentially and exhibits BP. However, consider the local cost function:

$$C_L = Tr\left[O_L V\left(\theta\right)|0\rangle\langle 0| V\left(\theta\right)^{\dagger}\right] \tag{6}$$

where $O_L = I = \frac{1}{n}\sum_{j=1}^{n} |0\rangle\langle 0|_j \otimes I_{-j}$, the local cost function is converted to $C_L = 1 - \frac{1}{n}\prod_{j=1}^{n} \cos^2 \frac{\theta^j}{2}$. The variance of its gradient is $Var\left[\frac{\partial C_L}{\partial \theta^j}\right] = \frac{1}{8n^2}$, which vanishes polynomially with n and exhibits no BP.

Marrero et al. analyzed the barren plateau phenomenon in terms of quantum circuit structure and concluded that excessive entanglement between visible and hidden units leads to BP [33]. Standard inequalities for the quantum relative entropy are:

$$\frac{1}{2}\left\|Tr_h\left(\rho\right) - I_v/D_v\right\|_1^2 \leq S\left[Tr_h\left(\rho\right)||I_v/D_v\right] = -S\left[Tr_h\left(\rho\right)\right] + \log\ \left(D_v\right) \tag{7}$$

Then, from the von Neumann trace inequality, we have

$$\left|Tr\left[\left(O_{Obj} \otimes I_h\right)\left(\rho - I_v/D_v\right)\right]\right| = X \tag{8}$$

$$\left\|O_{Obj}\right\| \infty \left\|Tr_h\left(\rho\right) - I_v/D_v\right\|_1 = Y \tag{9}$$

$$\left\|O_{Obj}\right\| \infty \sqrt{2\left\{\log D_v - S\left[Tr_h\left(\rho\right)\right]\right\}} = Z \tag{10}$$

$$X \leq Y \leq Z \tag{11}$$

If ρ satisfies an area-law scaling, then $S\left[Tr_h\left(\rho\right)\right] \in \Theta\left(n_v^{(1)}\right)$, from which the claimed result for the area-law scaling immediately follows. If, instead, we assume that ρ obeys a volume law, then $|S\left[Tr_h\left(\rho\right)\right]| - \log D_v \in \Theta\left(2^{n_v - n_h}\right)$ This shows that if our quantum neural network outputs state that satisfy a volume law, then asymptotically the predictions of the neural network would be no better than random guessing. In contrast, quantum neural networks will not necessarily observe this problem if the entanglement entropy is characteristic of an area-law scaling unless the number of hidden units in the first layer becomes much larger than the number of visible units. We therefore see that uncontrolled entanglement, such as that yielded by volume laws, can be catastrophic for a quantum variant of deep-neural networks that requires a number of hidden units much larger than the visible units For example, it was shown that the training of QNN satisfying the volume law of entanglement entropy is difficult because the expectation value of the objective function vanishes exponentially with the number of hidden units. Patti et al. sought to alleviate the BP measures by establishing the entanglement metric through random entanglement [34]. Quantum entanglement is a powerful tool for quantum computing, and it is necessary to grasp the balance between entanglement resources.

In the NISQ era, one of the advantages of the VQA algorithm is that the noise can be relieved. However, noise can also lead to BP, called noise-induced barren plateau (NIBPS) [35]. According to the proposition in reference, the measurement noise has the following formula.

$$\left|\widetilde{C} - \frac{1}{2^n} Tr\left[O\right]\right| \leq q_M^w G\left(n\right) \left\|\rho - \frac{1}{2^n}\right\|_1 \quad and \left|\partial_{lm}\tilde{C}\right| \leq q_M^w F\left(n\right) \tag{12}$$

It shows that in the presence of measurement noise there is an additional contribution from the locality of the measurement operator. For the local Pauli noise that acts on the entire quantum circuit, when the depth of ansatz grows linearly with N, the gradient vanishes exponentially in the number of qubits N. However, unlike the BP, the vanishing of the gradient is not the disappearance of random probability but is associated with the system size. Therefore, the solution strategy of the NIBP and BP cannot be applied to each other and requires quantitative guidance.

2.2 Method of representation of barren plateau

Zhao et al. proposed a general scheme to analyze the BP phenomenon in QNN using ZX-calculus [42]. ZX-calculus is a graphical language for reasoning about quantum computation, unlike circuit diagrams, where zx-algorithms display the properties, entanglement states, and protocols of a circuit in its entirety through graphs [43]. They consider a specific PQC given the expectation and gradient variance of the Hamiltonian quantum H-analysis cost function as a ZX-graph. A tree tensor network is shown in Fig. 3.

The gradient variance is calculated through the ZX-calculus as shown in Fig. 4.

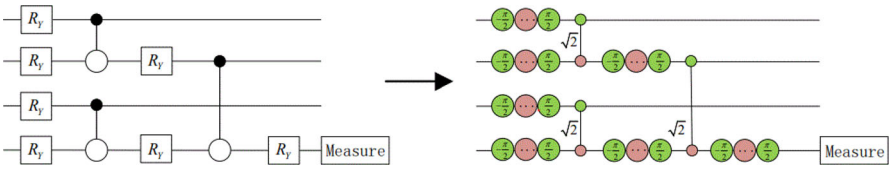

Fig. 3 Schematic diagram of a tree tensor network ZX-calculus [42]

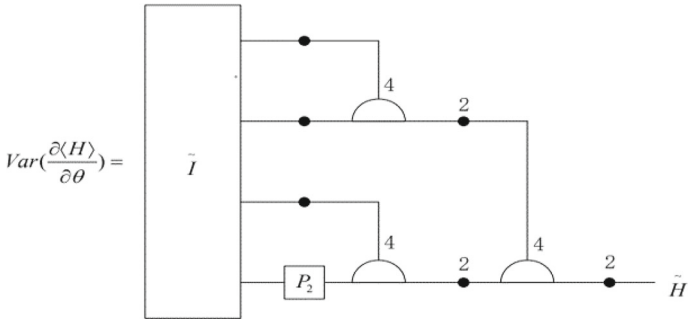

Fig. 4 Schematic diagram of a tree tensor network ZX-calculus [42]

In recent years, researchers have mainly demonstrated the BP phenomenon visually with cost landscapes as well as gradient variances, and the ZX-calculus has not been widely discussed due to the complexity of the circuit.

3 Research progress in solving barren plateau

In recent years, there are many solutions to BP. In this section, we will classify the solutions of the last 5 years according to the constructor of the QNN.

3.1 Solution based on quantum embedding

The ability of quantum algorithms to perform efficient computations is due in large part to the ability to grow the number of manipulations of a system in Hilbert space at most polynomially with the size of the system. Quantum embedding encodes classical states into quantum states and usually refers to quantum feature mapping, i.e., mapping data into quantum Hilbert space. Quantum embedded circuits are usually neither optimized nor trained and have multiple encoding methods, and the different encoding methods affect the computational power of QNN [44, 45].

Shin et al. proposed an exponential data coding strategy QFFLM [46], based on the above FFLM [47–49] computed by a variational circuit. Using the quantum supervision (QSL) model, in the case of a given small number of encoding gates or qubits, the encoding gates can generate an exponentially large Fourier-feature linear model. Use the QSL model to efficiently reduce the dimension of the input quantum state to alleviate the BP phenomenon caused by finding the minimum loss function.

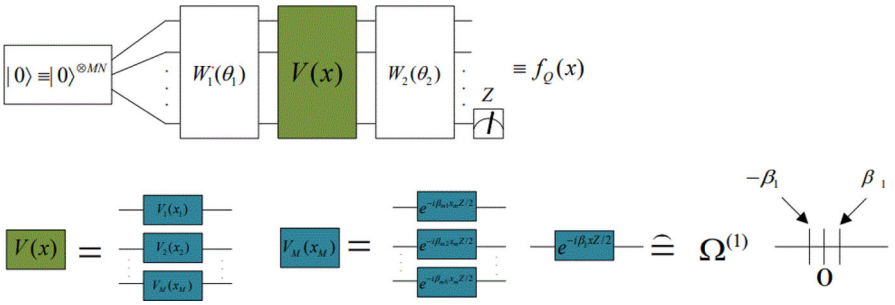

Fig. 5 Schematic diagram of QSL model used to express fQ(x) [46]

As shown in Fig. 5, the QSL model contains training unitary $W_l\theta_l$ and data encoding unitary $V(x)$. Each $W_l\theta_l$ contains L single-qubit layers and complete adjacent CNOT gates. $V(x)$ encodes each element x_m of a training datum x onto an N-qubit Pauli-Z gate with prechosen weights $\beta_{m1}, \beta_{m2}, ..., \beta_{mN}$. $f_Q(x)$ is a finite M-variate Fourier series:

$$f_Q(x) = \langle 0| U^{\dagger}_{x;\theta} Z U_{x;\theta} |0\rangle \equiv \sum n_{1\in\Omega_1} ... \sum n_{m\in\Omega_M} C_{n1,n2...nM} e^{-in\cdot x} \tag{13}$$

This exponential encoding strategy has the advantage of using a small number of qubits to represent the general function of the Fourier series, which can be used to obtain an arbitrary Fourier spectrum by adjusting the training data encoding weights, linking the classical Fourier characteristic model with quantum computing and avoiding the BP phenomenon arising from the process of minimizing the loss function. When computational resources are limited, it is also possible to express functions that lie outside the classical expressible region, highlighting its hidden potential. However, it is still limited to the Fourier coverage and has some restrictions on the circuit and training parameters, and is not widely used in circuit coding schemes.

Selvarajan et al. built a variational autoencoder with the data set as circuit input and output a state with half the number of qubits to reduce the dimensionality of the state in the high-dimensional Hilbert space [50], as shown in Fig. 6.

The variational autoencoder consists of two stages. In the first stage, the circuit shows D repeating layers encoding circuit $U(\vec{\theta})$, which consists of independent single-qubit Ry gates and a set of adjacent CNOT gates. Thus, it obtains $U(\vec{\theta})|\psi\rangle = |\varphi\rangle \otimes |\eta\rangle$, φ is the first subsystem A, which will be indexed later with an initial state $|0\rangle$, and η is the second subsystem B, which will be indexed later with an initial state $|1\rangle$. The unitary $U(\vec{\theta})$ decomposes $|\Psi_i\rangle_{A.B}$ into $|\varphi(\theta)_i\rangle_A \otimes |\eta(\theta)_i\rangle_B$. To obtain such a tensor product structure, the entropy on subsystem A or B needs to be minimized to zero, so that the cost function can be optimized. In the second stage, it uses an additional ancillary qubit, and a CSWAP gate is acted on the qubits of systems A and B. Thus, we get $|0\rangle\ |\psi\rangle + |1\rangle\ |\eta\rangle$. This encoding scheme transforms the input state containing 2n qubits into n+1 qubit states. The output dimension can be reduced repeatedly as long as the output states allow.

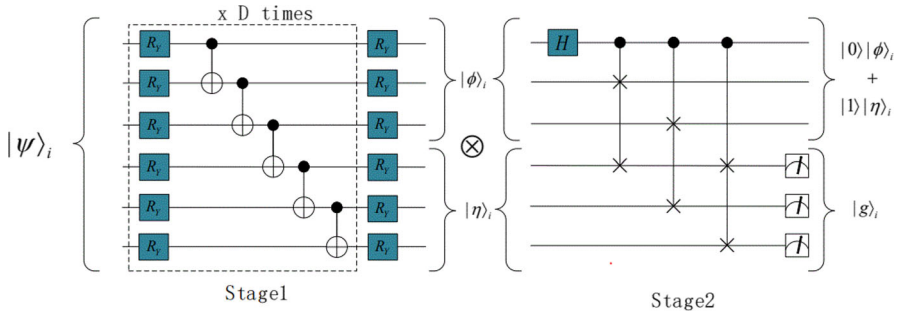

Fig. 6 Schematic diagram of variational autoencoder. The first subsystem optimizes over the dataset to generate equivalent states with the subsystem tensor product structure. The second subsystem performs measurements to assign relative phase and amplitude, preserving the characteristics of the starting state [50]

Although the models used by Shin and Selvarajan are different, their thoughts are derived from the dimension of dimension reduction in classical machine learning algorithms, and dimension reduction plays a key role in the training of machine learning. Therefore, effective coding and dimensionality reduction in QNN are likely to provide great help in solving effective training in high-dimensional Hilbert spaces, and it is worthwhile to continue to investigate deeply this kind of BP relief in the coding stage.

3.2 Solution of ANSATZ

3.2.1 Solution of ansatz parameters

Like the weight matrix in a classical neural network, the parameters in a QNN exist mainly in the quantum circuit, which represents a parameterizable unitary matrix $U(\theta)$. In general, we can arbitrarily adjust the parameters in a quantum circuit. In practice, the parameters θ are encoded as single-qubit rotation gates and double-qubit CNOT, CZ gates.

For a QNN, every small change in the parameters θ in the PQC has an impact on the cost function. For this, Grant et al. developed an initialization strategy [51], where some initialization values are randomly selected and then the remaining values are chosen so that the initialization parameters are a fixed unitary matrix. This makes the circuit a sequence of shallow blocks that effectively reduces the effective depth when the first parameter update is performed. Even if a single parameter is changed in the next block, the rest of the blocks still work as marked. This approach limits the range of initial parameter values mainly by changing the values of the parameters at the first optimization. Although it effectively alleviates the BP, it is very limited and does not guarantee that the selection point of the first parameter value is good enough, and optimizing the initialization parameters only once and only once does not guarantee the starting point for the next round of optimization. Zhang et al. propose a Gaussian initialization strategy for both global and local observable cases, which theoretically guarantees the trainability of the circuit [52] and is more applicable to the VQA.

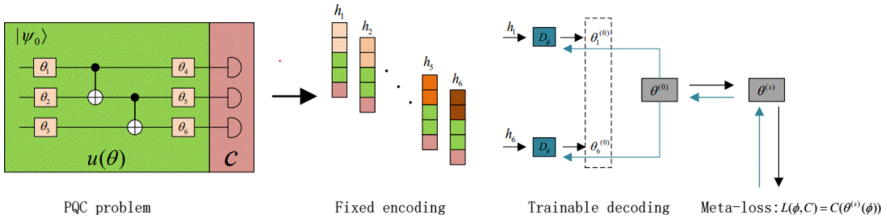

Fig. 7 Schematic diagram of FILP process [54]

Skolik et al. argue that this initialization method still has a large number of training parameters throughout the training process, and they propose to use a layered learning approach, where each circuit component is added to the training in turn, increasing the training parameters in turn and thus increasing its representational power [53]. In the first stage, a small number of starting layers of circuits are trained, in which all parameters are initialized to zero. After determining the number of layers to be trained, another set of layers is added, freezing the parameters of the previous layers. An iteration is defined as an update of all trainable parameters. In the second stage, the pre-trained circuit is partitioned into larger successive layers at a time, such as one-fourth or half of the circuit layers, alternating the subset of parameters trained until convergence of the objective function is reached. This layered learning approach is effective in reducing the BP phenomenon, because, in the first stage, the circuit maintains fewer sufficiently non-random initialization parameters, while the random ones are present in shallower subcircuits. However, this structure also leads to a deeper circuit depth, resulting in a waste of quantum gate resources.

Sauvage et al. proposed a framework FLIP based on the idea of meta-learning: a flexible initializer for arbitrarily sized PQC [54]. FLIP is highly flexible and can be applied to circuits of different sizes (including circuits with different qubits, different circuit widths, depths and number of variational parameters) to solve parameter initialization tasks. The core flow of FLIP is shown in Fig. 7. For a PQC parameter update problem, a, it is composed of an ansatz $U\ (\theta)$ and an objective C which can be estimated through repeated measurements on the output state $|\Psi\ (\theta)\rangle\ =\ U\ (\theta)\ |\Psi_0\rangle$. b, The problem is mapped to a set of initialization parameters $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$, which are encode into vectors $(h_1, h_2, h_3, h_4, h_5, h_6)$ containing information about the parameters themselves, the overall circuit, and optionally the objective. And each encoding is fixed, uniquely representing each parameter. These encodings are then decoded by a neural network, D_φ with weights φ, outputting a single value θ_k^0 per encoding h_k. c, these parameters $\theta^{(0)}$ are used as the starting point of a gradient descent optimization. During the training of FLIP, they are used to minimize the meta-loss function $L\ (\varphi, c)$, and the weights φ of the decoder are updated by back-propagated.

LIU et al. propose a migration learning-inspired method for optimizing initialization parameters to alleviate the BP [55]. PQC and optimal parameters of a small-scale task are transferred to a large-scale task. As shown in Fig. 8, it is divided into network transmission and structured transmission. For network transfer, $\theta*$ denotes the trained optimal parameters, θr is the random initialization parameter, and the small-scale task base circuits $U\theta_{1,n}$. Structure transfer, where the structure is transferred from 2 qubits

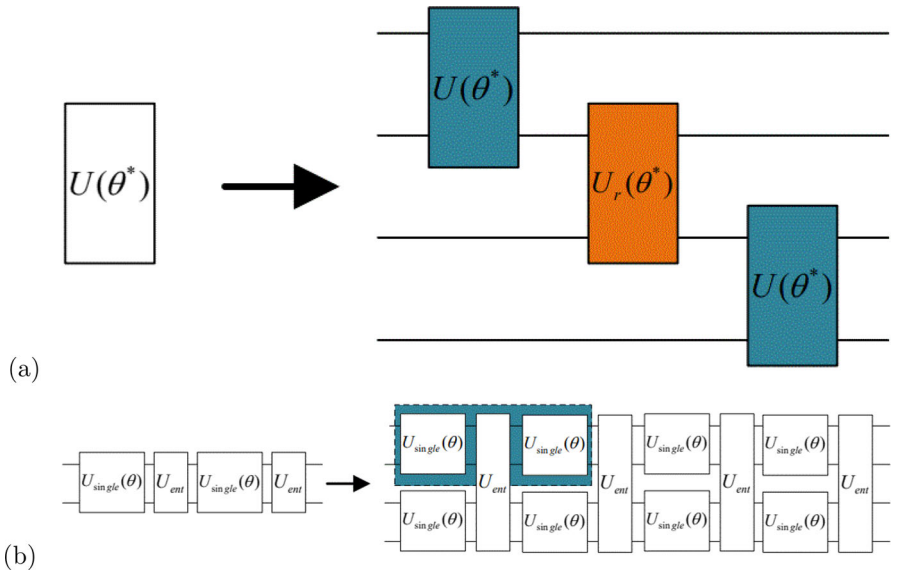

Fig. 8 Schematic diagram of transfer methods [55]. **a** Network transfer. **b** Structure transfer

to 4 qubits, is specified by the dashed line for each transfer block, either copied over from the small-scale task or randomly initialized underlying circuits. Although this approach has been shown to be effective in alleviating BP, small-scale tasks must be associated with large-scale tasks, and it is determined that optimal training of small-scale tasks is not necessarily very simple, and the complete training cost of small-scale and large-scale tasks remains to be explored.

We found roughly the same starting point in parametrically solving the BP problem as in classical neural networks solving the gradient disappearance or gradient explosion problem. We find roughly the same starting point in parametrically solving the BP problem as in classical neural networks solving the gradient disappearance or gradient explosion problem. Therefore, Grant et al. took the lead in fixing a small number of parameters in the initial circuit. Zhang et al. use Gaussian strategy. Skolik et al. create the starting layer and set all the parameters to zero to avoid the parameters initialized on the BP. Through the layered circuit increase, limit training and randomization on a small number of circuits. This is also similar to the layered pre-training of the classical neural network [56].

The FILP framework proposed by Sauvage et al. differs from the meta-learning parameter initialization method of classical neural networks, most notably in its higher flexibility, and while demonstrating significant advantages in the BP, the QAOA maximum cut problem, and finding model ground states, it does not analyze whether encoding–decoding parameters from PQC into classical neural networks have a high cost.

It can be seen that, in terms of parametric solutions, the use of narrowing the initialization parameters, reducing the total number of quantum circuits to be run, and adopting flexible initialization strategies to avoid BP has been quite successful.

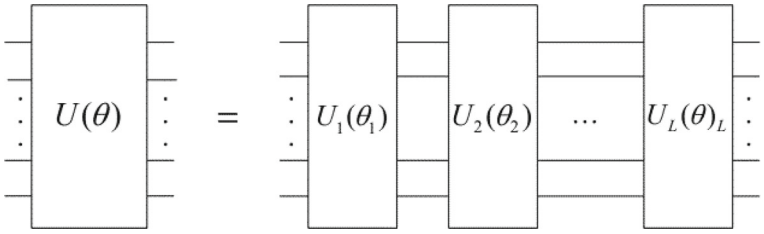

Fig. 9 Schematic diagram of an ansatz [7]. The unitary $U(\theta)$ is generally expressed as a product of L unitaries $U_i(\theta_i)$, $U(\theta) = U_L(\theta_L)...U_2(\theta_2)U_1(\theta_1)$. Each unitary $U(\theta)$ can in turn be decomposed into a single-qubit rotation gates and double-qubit gates. The parameters θ are encoded in the unitary $U(\theta)$, $U_l(\theta_l)\prod_m e^{-i\theta_m H_m} W_m$, where W_m is a unparameterized gates and H_m is a self-adjoint operator

3.2.2 Solution of ansatz structure

Another important component of the VQA is ansatz. As shown in Fig. 9, Ansatz is defined as the unitary circuit $U(\theta)$ acting on the input states of N qubits. In general, there is no fixed structure of Ansatz, the exact structure depends on the given task and how the parameters are trained to minimize the cost function. Therefore, a well-designed ansatz circuit structure affects the performance of the variational quantum algorithm.

Cenk et al. proposed a new method [57] to avoid BP by splitting the ansatz of N qubits into multiple sets of O(logN) qubits through classical splitting. Classical splitting splits a large number of qubits into a circuit of many fewer qubits, significantly reducing the number of two-qubit gates. The ansatz $U(\theta) = \prod_{l=1}^{L} U_l(\theta_l) W_l$ can be classically split into:

$$U(\theta) = \prod_{l=1}^{L} \otimes_{i=1}^{k} U_l^i\left(\theta_l^i\right) = \otimes_{i=1}^{k} \prod_{l=1}^{L} U_l^i\left(\theta_l^i\right) = \otimes_{i=1}^{k} U^i\left(\theta^i\right) \tag{14}$$

The ansatz after the classical split is shown in Fig. 10. The cost function becomes:

$$C(\theta) = \sum_{i=1}^{k} Tr\left[O_i U^i\left(\theta^i\right) \rho_i U^{i\dagger}\left(\theta^i\right)\right] \tag{15}$$

where the costs of each classically split circuit are independent of each other. The gradient of j^{th} parameter of the j^{th} ansatz is:

$$\partial_{i,j} C(\theta) = \partial_{i,j}\left(Tr\left[O_i U^i\left(\theta^i\right) \rho_i U^{i\dagger}\left(\theta^i\right)\right]\right) \tag{16}$$

The variance of the gradient is:

$$Var\left[\partial_j C(\theta)\right] \approx O\left(\frac{1}{2^{6m}}\right) = O\left(\frac{1}{N^{6\beta \log_\gamma 2}}\right) \tag{17}$$

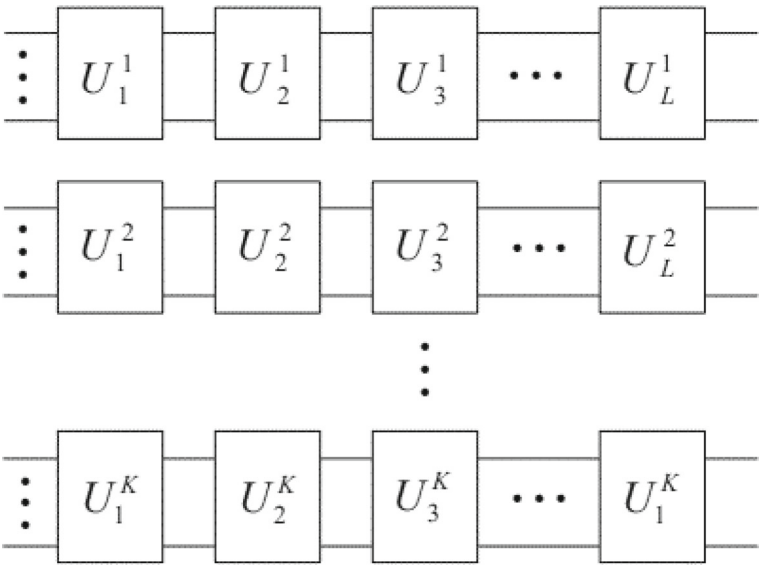

Fig. 10 Schematic diagram of a classical split ansatz [57]. A L layers of the ansatz are decomposed into $k = \frac{N}{m}$ many qubit U blocks

Thus, the variance of the classically split ansatz satisfies $O\left(poly\,(N)^{-1}\right)$ and does not lead to the phenomenon of BP.

Du et al. proposed a quantum circuit architecture search (QAS) scheme to generate variable ansatz [58], which automatically finds the optimal ansatz based on a given learning task to achieve the optimal performance of quantum circuits. QAS is divided into four steps, as shown in Fig. 11: (1) supernet setup, (2) optimize supernet, (3) ranking, and (4) retrain the searched ansatz.

In step 1, QAS sets up a supernet to create a set of quantum gates pools for target search and parameterizes the quantum gates pools by a weight sharing strategy. Single-qubit gates are represented using hexagons, and two-qubit gates are represented using rectangles, with U_x representing the data encoding layer.

In step 2, QAS minimizes the objective function, draws single quantum gates from a pool of quantum gates, and optimizes their trainable parameters. The correlation parameters between different sequences are minimized by a weight sharing strategy.

In step 3, QAS compares the performance between 1 and K sequences based on the training parameters θ and the objective function. The optimal ansatz is selected as the most output.

In step 4, QAS is retrained using the optimal ansatz and parameters according to the objective function.

Two key elements of QAS are the supernet sharing strategy and weight sharing strategy, which reduce unnecessary multiple quantum gates and lower the depth and entanglement of the circuit. QAS differs from [59] fixed structure optimization scheme by maintaining a shallow circuit depth while satisfying sufficient expressiveness and reducing computational overhead. QAS removes unnecessary quantum gates to reduce

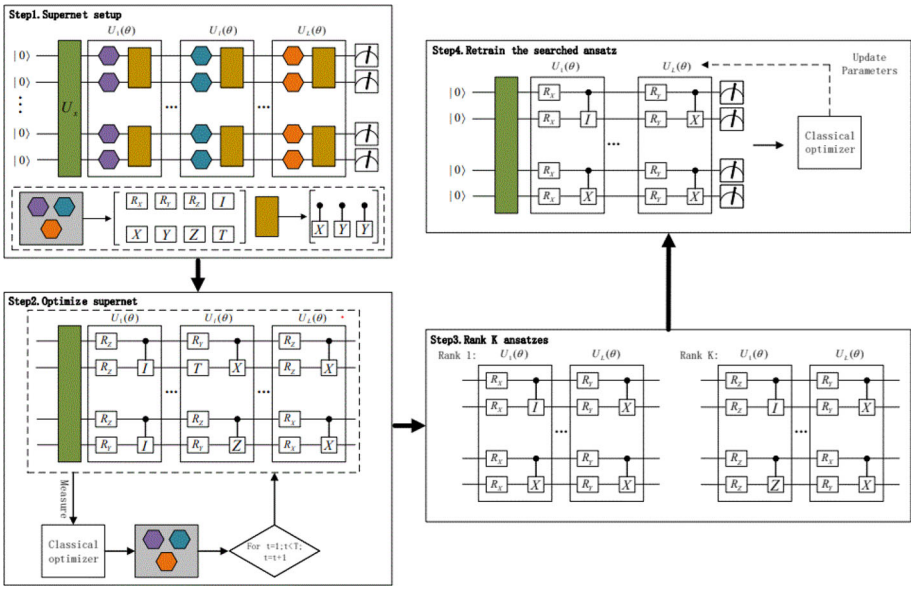

Fig. 11 Schematic diagram of the quantum architecture search scheme (QAS) [58]

entanglement in the circuit while maintaining the trainability of the circuit and alleviating the BP.

Wiersema et al. focused on a particular type of quantum circuit, the Hamiltonian variational ansatz (HVA). It was shown that its circuit structure exists as a variational quantum characteristic solver with slight or no BP [60]. [61] which does actually rigorously characterize parameter conditions such that the HVA does avoid barren plateaus. HVA [62] is a QAOA [8] and adiabatic calculations based on [62]. As shown in Fig. 12, many Hamiltonian terms are included.

Experiments on the entanglement and gradient variance of HVA proved its good structural properties to relieve BP. Mele [63] et al. further improved on HVA by iterating the difference scheme [64] to obtain a smooth optimal solution, near which the landscape is not affected by the gradient thus relieve BP.

Li et al. proposed a new hybrid quantum-classical framework called variational shadow quantum learning(VSQL) [65], as shown in Fig. 13. First, a parametric shaded quantum circuit $U\ (\theta)$ is used to extract classical features in a convolutional manner. This shaded feature approach labels only local sub-controls in the quantum Hilbert space instead of the whole Hilbert space. Secondly, the shadow features of the input data ρ_{in} are computed from the measured $pauli - (X \otimes X)$ expectation values. The resulting shadow features are then fed into a classical fully connected neural network for post-processing. The value of the label $\hat{y}$ between 0 and 1 is determined by an activation function. This approach reduces the number of parameters and facilitates the training of quantum circuits. They found that VSQL can avoid BP by analyzing the loss landscape. Sack et al. also used the classical shading approach [66], and they proposed a concept of weak barren plateaus (WBP) based on the entropy of the local

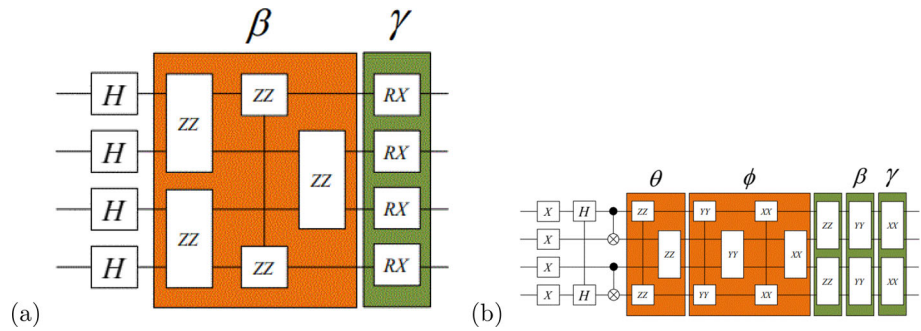

Fig. 12 Schematic diagram of HVA quantum circuit [60]. **a** HVA quantum circuit for TFIM, **b** HVA quantum circuit for the XXZ model

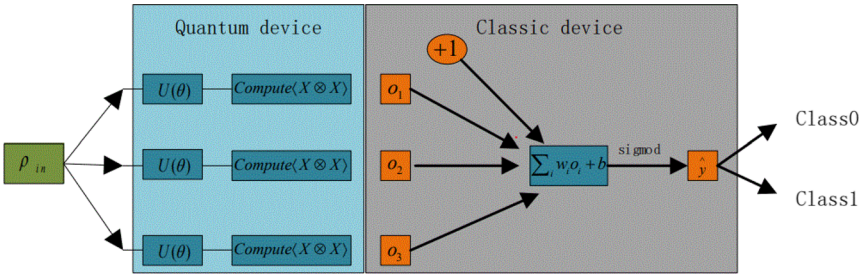

Fig. 13 Schematic diagram of variational shadow quantum learning (VSQL) for binary classification [65]

simplified density matrix. The presence of WBP can be quantified using classical shading that avoids BP by avoiding local WBP.

Unique ansatz can be designed to meet the needs of the respective tasks. It also expresses that the strategy to solve the BP cannot be extended to general-purpose circuits. Marrero, Patti et al. have also shown that the structure of the circuit is one of the main factors that lead to the BP phenomenon. Fortunately, Pesah et al. rigorously proved that there is no BP in quantum convolutional neural networks [67], and Zhao et al. also proved that there is no BP in quantum convolutional neural networks and tree tensor networks. This suggests that finding alternative architectures is also a strategy for QNN. However, only these two architectures have been proven, and the next quantum graph convolutional neural networks, and quantum generative adversarial networks, are all good research directions that deserve further exploration by researchers.

3.3 Solution of optimize

One of the main differences between VQA and classical machine learning algorithms is that while classical machine learning models are neural network models running on classical computers, VQA is quantum circuits running on quantum computers. As mentioned earlier, we have finished analyzing the implications and solutions arising in

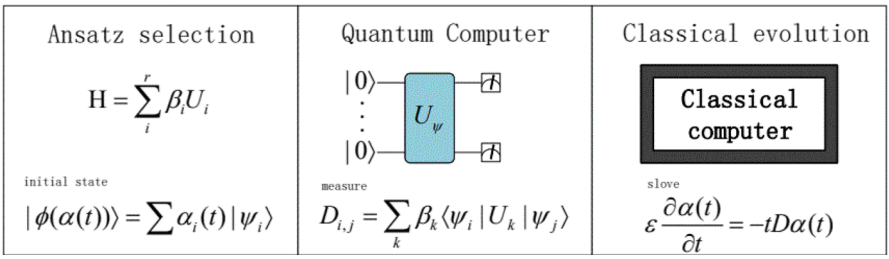

Fig. 14 Schematic diagram of the quantum assisted simulator (QAS) algorithm [68]

terms of quantum circuits. Next, we will analyze the solutions in terms of the optimal parametric search for VQA.

The VQA is parameter optimized on a classical computer so that many classical machine learning optimization algorithms can be applied. VQA typically use a gradient descent algorithm to optimize parameters, minimize the loss function, and feed the parameters back to the quantum circuit until the loss function or the number of iterations satisfies the stopping condition. Given this process, two main solutions are currently proposed, one for the parameter feedback process and one for the gradient descent method.

For the parametric feedback process, Bharti and Haug et al. tried to remove the classical quantum feedback loop and provided a new algorithm, the Quantum Assisted Simulator, by which the BP was avoided by this structure [68]. As shown in Fig. 14, in step 1, ansatz selection, the Hamiltonian H is a combination of r unitaries $H = \sum_i^r \beta_i U_i$, and the ansatz is a linear combination of cumulative K-moment states $|\phi(\alpha(t))\rangle = \sum_{|\psi_i\rangle \in CS_k} \alpha i(t)|\psi i\rangle$. In step 2, calculation of the overlap matrices on a quantum computer, $D_{i,j} = \sum \beta_k K \langle\psi_i| U_k |\psi_j\rangle$. In step 3, solving the differential equation $\varepsilon \frac{\partial \alpha(t)}{\partial t} = -tD\alpha(t)$, thus update of the parameters becomes $\alpha_j(t+\delta t) = \alpha_j(t) + \alpha\delta t$.

This method of avoiding quantum circuit feedback does not require the calculation of gradients, the quantum state of the ansatz is fixed, and only the variational parameters are updated. The generation of BP is solved from the root. However, it also changes the original purpose of the VQA, which is a good solution, but not the best one.

Zhang et al. proposed a gradient-free quantum learning framework Quark for classification tasks to optimize quantum models with quantum optimization (QMOO) [69]. As shown in Fig. 15, it consists of three parts: the first part, Ψ_0 preparation, which initializes R_w, R_D, R_O. Initialize the weights register R_w using base encoding on each qubit. Data register R_D, using K parallel data sets, entangles the training data set and the model weights in parallel. Auxiliary register R_O stores intermediate results and model outputs, by forwarding the quantum model architecture. In the second part, amplitude amplification is performed using the Grover algorithm, where each Grover iteration amplifies the probability of the optimal weights to the maximum. In the third part, the weights are measured, and the parallel data set improves the probability of the optimal weight in each measurement.

This framework designed by the quantum optimization approach firstly uses gradient-free optimization in terms of weight optimization, which is performed by weight probability height to avoid BP, and secondly uses parallel data sets that can be scaled exponentially to further optimize the algorithm. It cannot be equated to gradient-free optimization algorithms, and gradient-free optimizers still have BP [70]. On the other hand, Suzuki et al. applied the normalized gradient descent method to VQA [71], which can overcome the gradient disappearance phenomenon to avoid the BP. Kieferova et al. proposed a training algorithm that minimizes the second-order maximum Rényi scatter [72] to circumvent the BP by using an unbounded loss function. The loss function of existing quantum algorithms uses a linear bounded loss function, and when using the second-order maximum Raney scatter as the loss function, the minimization will reach the upper limit of the quantum simulation of the KL-divergence between two quantum states, resulting in the BP not being applicable. Although this method does not have BP, it has significant limitations and does not guarantee effective training.

This unique quantum-classical hybrid property of variational quantum algorithms is already well known to researchers. Therefore, whether it is avoiding parameter feedback or using other gradient descent algorithms, this is well documented in classical machine learning. It is certainly a wise shortcut to carry out quantum computer problem solving directions by drawing on the already developed solutions of classical machine learning algorithms for solving gradient vanishing or gradient explosion problems in neural networks. In addition, for many mitigation strategies for BPs, the strategy comes at a price of reducing the expressivity of the optimization. There are still significant challenges and opportunities in this field

3.4 Taxonomy

We analyze and compare the current solutions to the BP. As shown in Table 1, the "# of qubits" column shows the number of qubits required after any dimensional reduction is used, when the initial number of qubits is N.

At this stage, the problem of BP is addressed in three main aspects: 1. quantum embedding; 2. parameter selection and structural design of ansatz; and 3. optimization methods. In the selection of qubits, if not specifically mentioned, we all follow the given qubits N for comparison, and it can be seen that the reason for the change in the number of qubits is mainly due to quantum embedding, and the classical splitting scheme causes the change in the scale of qubits due to splitting. As for the circuit depth, it is not mentioned except for the exponential encoding scheme and the hierarchical learning scheme. The main reason is that the choice of circuit depth depends on the particular problem; for small-scale problems, the circuit depth can be chosen shallower and, conversely, deeper. Moreover, there is no uniform detection standard for each solution, so the circuit depth is not specifically stated in the scheme. For generality, the meta-learning FILP framework and the gradient-free quantum learning framework perform better. Most of the determination criteria for problem solving are based on the absence of gradient variance, and the rest of the proposed solutions, through numerical simulation experiments, such as hierarchical learning, indirectly

Table 1 Analysis of BP solutions

Solution			# of qubits	Circuit depth	Advantages	Limitation	Experimental verification	Experimental environment
Quantum embedding		Exponential data encoding [46]	$\log_3(2df+1)$	Polynomial depth	The exponential Fourier spectrum can be obtained by adjusting the training data weights, which has a quantum advantage over the classical Fourier model.	Restricted range, subject to Fourier coefficient coverage.	None	No noise
		Variational encoders [50]	$\frac{N}{2}$	Unknown	Half of the state dimension is represented in a high-dimensional Hilbert space.	Connectivity of large and small systems.	None	No noise
Ansatz	Ansatz parameters	Parameter initialization [51]	N	Unknown	Reduced circuit depth for the first parameter update	Only first-time optimization is guaranteed.	Gradient variance	No noise

Table 1 continued

Solution		# of qubits	Circuit depth	Advantages	Limitation	Experimental verification	Experimental environment
	Layerwise learning [53]	N	Deepen	The learning strategy is updated to a subset of the updated parameters each time to avoid large gradient amplitude during training.	Excessive quantum gates.	None	Noise
	FILP [54]	N	Unknown	It can be adapted to circuits of different sizes with a high degree of flexibility, using solutions that are more robust in terms of finding parameter space.	Cost of decoding-encoding	Gradient variance	Noise and no noise
	Transfer learning [55]	N	Unknown	High training efficiency	Small-scale tasks are associated with large-scale tasks	None	No noise

Table 1 continued

Solution		# of qubits	Circuit depth	Advantages	Limitation	Experimental verification	Experimental environment
Ansatz structure	Classical splitting [57]	O(log N)	Unknown	Reducing the exponential growth of the Hilbert space dimension has a faster convergence rate.	More parameters	Gradient variance	No noise
	HVA [62]	N	Unknown	The circuit has over-parameterization and still maintains good circuit characteristics.	Fixed circuit structure and low scalability	Gradient variance	No noise
	VSQL [64]	N	Unknown	Reducing the number of parameterized quantum gates is independent of the size of the problem.	Limited to classified tasks	Cost function	No noise
	Entanglement structure [33]	N	Unknown	The entanglement degree is quantified and the scheme is verified from the internal structure of the circuit.	Entanglement is inevitable	Cost function	No noise

Table 1 continued

Solution		# of qubits	Circuit depth	Advantages	Limitation	Experimental verification	Experimental environment
Optimize	Quantum assisted simulator [67]	N	Unknown	The quantum computer as the main computation is time superior to the iterative variational scheme.	May need to be re-optimized	None	No noise
	Gradient-free quantum learning framework-QUARK [68]	$\geq$ N	Unknown	The classical computer gradient optimization is not used and the problem of data set size can be disregarded.	It is difficult to achieve in small-scale quantum devices.	None	No noise

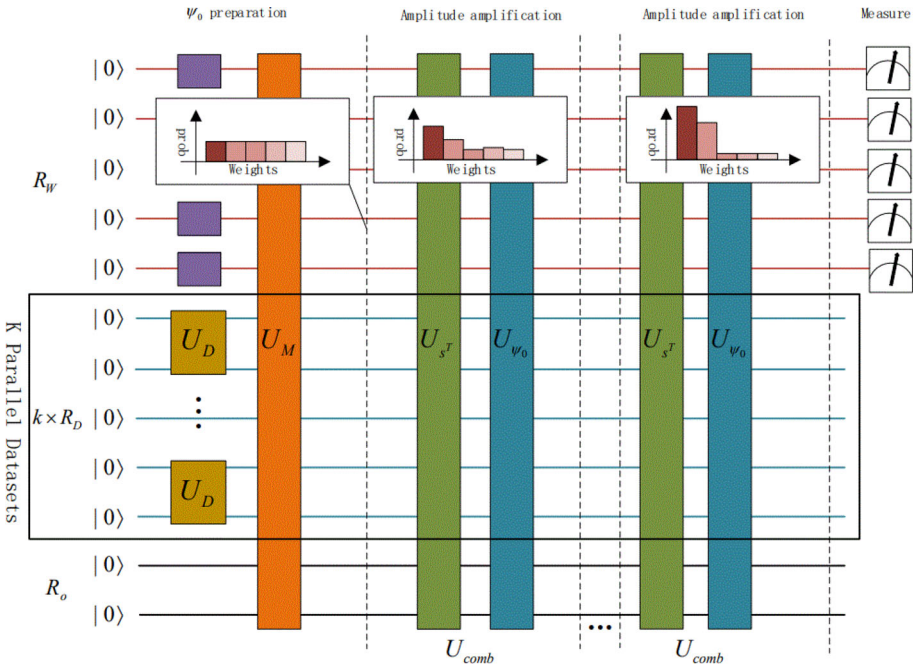

Fig. 15 Schematic diagram of the Quark optimization framework [69]

demonstrate the mitigation or absence of the BP by comparing the accuracy improvement of a classification task.

Different schemes choose different entry points, but the final directions are extremely consistent. For example, variational self-encoder, meta-learning FILP framework, and classical splitting ansatz, starting from three directions, all eventually reduce the dimensionality of the target states in Hilbert space. Drawing on the ideas of hierarchical learning and classical splitting, grouping the dataset is seen as a feasible solution. Parameter initialization, migration learning parameter initialization, Gaussian parameter initialization, and other parameter initialization are chosen for initialization selection by different algorithms, and it is also feasible to choose other classical optimization algorithms in comparison with classical neural network optimization. Variable molecular encoder and variable fractional shadow quantum circuit are likewise an extension of classical machine learning on quantum.

Finally, we also summarize the experimental environment of each solution, except for hierarchical learning and FILP framework involves with noise, the rest are in an ideal noise-free environment, and in NISQ devices, quantum circuits are unable to ignore the effect of noise, which means that most of the solutions are still based on theory and are not applied to reality.

4 Trends

Over the past decade or so, quantum computing has moved from theory to reality, with increasingly mature quantum technologies providing opportunities for the implementation of quantum algorithms, and quantum computers developed by IBM and other companies enabling researchers to manipulate dozens of qubits to implement simple quantum circuits. More and more researchers are also getting involved in the BP discussion, the BP has so far received a variety of solutions, we have only reviewed the current phase of research solutions that are mainly aimed at alleviating the BP, and many solutions involve the BP but are not the real target.

Powerful solutions at this stage can be summarized as either reducing the dimensionality of the Hilbert space or avoiding gradient optimization. Solutions concerning parameters, although many are proposed, do not always guarantee that the parameters are chosen to satisfy trainability. As for the design of the ansatz structure, entanglement is one of the advantages of quantum computing, and a general solution cannot be proposed. Unless quantum convolutional neural networks, which are unique to QNN, have been shown not to have BP. Currently, research addressing the barren plateau problem in measurement directions within quantum computing has not achieved significant results. However, a deeper consideration of the cost function and its local variants' potential impact on the measurement process, along with the recognition that advanced network architectures such as quantum convolutional neural networks and tree tensor networks appear to be unaffected by the barren plateau problem, and guarantee the uniqueness of output measurements, offers a new perspective for research. Based on this, quantum generative adversarial networks and quantum tensor networks may become focal points of future research, holding the promise of being the next generation of efficient model architectures.

State preparation to the final classical computer optimization, appropriate solutions can be found at each stage. However, there is a lack of universal, scalable solutions. Although McClean and Cerezo et al. explained the BP phenomenon, they showed that the BP was related to the cost function. However, the phenomenon of BP in noisy quantum circuits is not discussed. Recent solutions seem to ignore the effects of noise, which has led to impressive results for small-scale quantum systems, but no breakthrough for large-scale quantum computers.

The creation of BP leads to the untrainability of QNN, and the design of QNN will be a trade-off between BP and trainability to achieve quantum dominance. Other interesting studies include a detailed discussion of how quantum noise will affect the BP, another phenomenon called laziness in QNN in addition to the BP phenomenon, and the effect of over-parameterization in quantum circuits. On the other hand, as QNN forms a strict system framework, the quantum hybrid classical algorithm based on the VQA algorithm is transformed into a quantum–quantum true quantum circuit model and its quantum learning algorithm. All these imply that the research on QNN has a long way to go and the BP will only be a short-lived challenge.

Data Availability All data, models, and code generated and used during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Preskill, J.: Simulating quantum field theory with a quantum computer (2018)
2. Endo, S., Kurata, I., Nakagawa, Y.O.: Calculation of the Green's function on near-term quantum computers (2019)
3. Cao, Y., Romero, J., Olson, J.P., Degroote, M., Aspuru-Guzik, A.: Quantum chemistry in the age of quantum computing (2018)
4. Outeiral, C., Strahm, M., Shi, J., Morris, G.M., Benjamin, S.C., Deane, C.M.: The prospects of quantum computing in computational molecular biology. Wiley Interdiscip. Rev. Comput. Mol. Sci. **11** (2021)
5. Thanasilp, S., Wang, S., Nghiem, N.A., Coles, P.J., Cerezo, M.: Subtleties in the trainability of quantum machine learning models (2021)
6. Preskill, J.: Quantum computing in the nisq era and beyond. Quantum **2**, 79 (2018)
7. Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S.C., Endo, S., Fujii, K., McClean, J.R., Mitarai, K., Yuan, X., Cincio, L., et al.: Variational quantum algorithms. Nat. Rev. Phys. **3**(9), 625–644 (2021)
8. Farhi, E., Goldstone, J., Gutmann, S.: A quantum approximate optimization algorithm. arXiv:1411.4028 (2014)
9. Lilienfeld, A.V.: Quantum machine learning. In APS March Meeting (2017)
10. Aspuru-Guzik, A., Peruzzo, A., Mcclean, J., O'Brien, J.L., Yung, M.H., Love, P.J., Shadbolt, P., Zhou, X.Q.: A variational eigenvalue solver on a quantum processor. Nat. Commun. **5**(1), 4213 (2013)
11. Higgott, O., Wang, D., Brierley, S.: Variational quantum computation of excited states. Quantum **3**, 156 (2019)
12. Aspuru-Guzik, A., Dutoi, A.D., Love, P.J., Head-Gordon, M.: Simulated quantum computation of molecular energies. Science **309**(5741), 1704–1707 (2006)
13. Mcardle, S., Jones, T., Endo, S., Li, Y., Yuan, X.: Variational ansatz-based quantum simulation of imaginary time evolution. npj Quantum Inf. **5**(1), (2019)
14. Yuan, X., Endo, S., Zhao, Q., Li, Y., Benjamin, S.C.: Theory of variational quantum simulation. Quantum **3**, 191 (2019)
15. Endo, S., Sun, J., Li, Y., Benjamin, S., Yuan, X.: Variational quantum simulation of general processes. Phys. Rev. Lett. **125**(1), 010501 (2018)
16. Bravo-Prieto, C., Larose, R., Cerezo, M., Subasi, Y., Cincio, L., Coles, P.: Variational quantum linear solver: a hybrid algorithm for linear systems. Am. Phys. Soc. (2020)
17. Huang, H.Y., Bharti, K., Rebentrost, P.: Near-term quantum algorithms for linear systems of equations (2019)
18. Lubasch, M., Joo, J.W., Moinier, P., Kiffner, M., Jaksch, D.: Variational quantum algorithms for nonlinear problems. Am. Phys. Soc. **101**(1), 010301 (2020)
19. Schuld, M., Bocharov, A., Svore, K., Wiebe, N.: Circuit-centric quantum classifiers. Phys. Rev. A **101**(3), 032308 (2018)
20. Romero, J., Olson, J.P., Aspuru-Guzik, A.: Quantum autoencoders for efficient compression of quantum data. Quantum Sci. Technol. **2**(4), 045001 (2016)
21. Benedetti, M., Garcia-Pintos, D., Perdomo, O., Leyton-Ortega, V., Nam, Y., Perdomo-Ortiz, A.: A generative modeling approach for benchmarking and training shallow quantum circuits. Springer Science and Business Media LLC (1) (2019)
22. Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Woerner, S.: The power of quantum neural networks. Nat. Comput. Sci. **1**(6), 403–409 (2021)
23. McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H.: Barren plateaus in quantum neural network training landscapes. Nat. Commun. **9**(1), 4812 (2018)
24. Harrow, A.W., Montanaro, A.: Quantum computational supremacy. Nature **549**(7671), 203–209 (2017)
25. Lund, A., Bremner, T.R.M.: Quantum sampling problems, bosonsampling, and quantum supremacy. npj Quantum Inf. **3**(1), 15 (2017)
26. Liu, J.G., Wang, L.: Differentiable learning of quantum circuit born machine. Phys. Rev. A **98**(6), 062324 (2018)

27. Farhi, E., Neven, H.: Classification with quantum neural networks on near term processors. arXiv:1802.06002 (2018)
28. Huggins, W.J., Patil, P., Mitchell, B., Whaley, K.B., Stoudenmire, M.: Towards quantum machine learning with tensor networks. Quantum Sci. Technol. **4**(2), 024001 (2018)
29. Grant, E., Benedetti, M., Cao, S., Hallam, A., Lockhart, J., Stojevic, V., Green, A.G., Severini, S.: Hierarchical quantum classifiers. npj Quantum Inf. **4**(1), 65 (2018)
30. Skolik, A., Jerbi, S., Dunjko, V.: Quantum agents in the gym: a variational quantum algorithm for deep q-learning (2021)
31. Holmes, Z., Sharma, K., Cerezo, M., Coles, P.J.: Connecting ansatz expressibility to gradient magnitudes and barren plateaus (2021)
32. Cerezo, M., Sone, A., Volkoff, T., Cincio, L., Coles, P.J.: Cost function dependent barren plateaus in shallow parametrized quantum circuits. Nat. Commun. **12**(1), 1791 (2020)
33. Marrero, C.O., Kieferova, M., Wiebe, N.: Entanglement-induced barren plateaus. PRX Quantum **2**(4), 040316 (2021)
34. Patti, T.L., Najafi, K., Gao, X., Yelin, S.F.: Entanglement devised barren plateau mitigation. Phys. Rev. Res. **3**(3), 033090 (2021)
35. Wang, S., Fontana, E., Cerezo, M., Sharma, K., Sone, A., Cincio, L., Coles, P.J.: Noise-induced barren plateaus in variational quantum algorithms. Nat. Commun. **12**(1), 6961 (2021)
36. Wu, A., Li, G., Ding, Y., Xie, Y.: Mitigating noise-induced gradient vanishing in variational quantum algorithm training. arXiv:2111.13209 (2021)
37. Tangpanitanon, J., Thanasilp, S., Dangniam, N., Lemonde, M.A., Angelakis, D.G.: Expressibility and trainability of parameterized analog quantum systems for machine learning applications. Phys. Rev. Res. **2**(4), 043364 (2020)
38. Holmes, Z., Sharma, K., Cerezo, M., Coles, P.J.: Connecting ansatz expressibility to gradient magnitudes and barren plateaus. PRX Quantum **3**(1), 010313 (2021)
39. Arrasmith, A., Holmes, Z., Cerezo, M., Coles, P.J.: Equivalence of quantum barren plateaus to cost concentration and narrow gorges. Quantum Sci. Technol. **7**(4), 045015 (2021)
40. Dankert, C., Cleve, R., Emerson, J., Livine, E.: Exact and approximate unitary 2-designs and their application to fidelity estimation. Phys. Rev. A **80**(1), 012304 (2009)
41. Harrow, A.W., Low, R.A.: Random quantum circuits are approximate 2-designs. Commun. Math. Phys. **291**(1), 257–302 (2009)
42. Zhao, C., Gao, X.S.: Analyzing the barren plateau phenomenon in training quantum neural network with the zx-calculus. Quantum **5**, 466 (2021)
43. John, V.D.W.: ZX-calculus for the working quantum computer scientist (2020)
44. Schuld, M., Killoran, N.: Quantum machine learning in feature Hilbert spaces. Phys. Rev. Lett. **122**(4), 040504 (2019)
45. Schuld, M., Sweke, R., Meyer, J.J.: Effect of data encoding on the expressive power of variational quantum-machine-learning models. Am. Phys. Soc. **103**(3), 032430 (2021)
46. Shin, S., Teo, Y.S., Jeong, H.: Exponential data encoding for quantum supervised learning. Phys. Rev. A **107**, 012422 (2023)
47. Biamonte, J.: Universal variational quantum computation. Phys. Rev. A **103**, L030401 (2021)
48. Cao, Y., Romero, J., Olson, J.P., Degroote, M., Aspuru-Guzik, A.: Quantum chemistry in the age of quantum computing (2018)
49. Endo, S., Cai, Z., Benjamin, S.C., Yuan, X.: Hybrid quantum-classical algorithms and quantum error mitigation. Journal of the Physical Society of Japan (3) (2021)
50. R.Selvarajan, M.Sajjan, T.Humble, S.Kais: Dimensionality reduction with variational encoders based on subsystem purification (2021)
51. Grant, E., Wossnig, L., Ostaszewski, M., Benedetti, M.: An initialization strategy for addressing barren plateaus in parametrized quantum circuits. arXiv e-prints (2019)
52. Zhang, K., Hsieh, M.-H., Liu, L., Tao, D.: Gaussian initializations help deep variational quantum circuits escape from the barren plateau (2022)
53. A.Skolik, J.McClean, M.Mohseni, P.Smagt, M.Leib: Layerwise learning for quantum neural networks (2021)
54. Sauvage, F., Sim, S., Kunitsa, A.A., Simon, W.A., Mauri, M., Perdomo-Ortiz, A.: Flip: A flexible initializer for arbitrarily-sized parametrized quantum circuits (2021)
55. Liu, H.Y., Chen, Z.Y., Sun, T.P., Wu, Y.C., Han, Y.J., Guo, G.P.: Mitigating barren plateaus with transfer-learning-inspired parameter initializations (2021)

56. Hettinger, C., Christensen, T., Ehlert, B., Humpherys, J., Jarvis, T., Wade, S.: Forward thinking: building and training neural networks one layer at a time (2017)
57. Tuysuz, C., Clemente, G., Crippa, A., Hartung, T., Kuhn, S., Jansen, K.: Classical splitting of parametrized quantum circuits (2022)
58. Yuxuan, D., Tao, H., Shan, Y., M.Hsieh, Dacheng, T.: Quantum circuit architecture search for variational quantum algorithms (2022)
59. Ostaszewski, M., Grant, E., Benedetti, M.: Structure optimization for parameterized quantum circuits. Quantum **5**, 391 (2021)
60. Wiersema, R., Zhou, C., De Sereville, Y., Carrasquilla, J.F., Kim, Y.B., Yuen, H.: Exploring entanglement and optimization within the hamiltonian variational ansatz (2020)
61. Park, C.-Y., Killoran, N.: Hamiltonian variational ansatz without barren plateaus. arXiv preprint arXiv:2302.08529 (2023)
62. Troyer, M., Wecker, D., Hastings, M.B.: Progress towards practical quantum variational algorithms. Phys. Rev. At. Mol. Opt. Phys. **92**(4), 042303 (2015)
63. Mele, A.A., Mbeng, G.B., Santoro, G.E., Collura, M., Torta, P.: Avoiding barren plateaus via transferability of smooth solutions in a Hamiltonian variational ansatz. Phys. Rev. A **106**(6), L060401 (2022)
64. Zhou, L., Wang, S.T., Choi, S., Pichler, H., Lukin, M.D.: Quantum approximate optimization algorithm: performance, mechanism, and implementation on near-term devices. Phys. Rev. X **10**(2), 021067 (2020)
65. Li, G., Song, Z., Wang, X.: Vsql:variational shadow quantum learning for classification (2020)
66. Sack, S.H., Medina, R.A., Michailidis, A.A., Kueng, R., Serbyn, M.: Avoiding barren plateaus using classical shadows. PRX Quantum **3**(2), 020365 (2022)
67. Pesah, A., Cerezo, M., Wang, S., Volkoff, T., Sornborger, A.T., Coles, P.J.: Absence of barren plateaus in quantum convolutional neural networks. Phys. Rev. X **11**(4), 041011 (2020)
68. Bharti, K., Haug, T.: Quantum-assisted simulator. Phys. Rev. A **104**(4), 042418 (2021)
69. Zhang, Z., Chen, Z., Huang, H., Jia, Z.: Quark: a gradient-free quantum learning framework for classification tasks (2022)
70. Arrasmith, A., Cerezo, M., Czarnik, P., Cincio, L., Coles, P.J.: Effect of barren plateaus on gradient-free optimization. Quantum (2020)
71. Suzuki, Y., Yano, H., Raymond, H., Yamamoto, N.: Normalized gradient descent for variational quantum algorithms (2021)
72. Kieferova, M., Carlos, O.M., Wiebe, N.: Quantum generative training using r'enyi divergences (2021)