

ARTICLE

OPEN



Implementation of quantum imaginary-time evolution method on NISQ devices by introducing nonlocal approximation

Hirofumi Nishi ^{1,2}✉, Taichi Kosugi ^{1,2} and Yu-ichiro Matsushita ^{1,2}

The imaginary-time evolution method is a well-known approach used for obtaining the ground state in quantum many-body problems on a classical computer. A recently proposed quantum imaginary-time evolution method (QITE) faces problems of deep circuit depth and difficulty in the implementation on noisy intermediate-scale quantum (NISQ) devices. In this study, a nonlocal approximation is developed to tackle this difficulty. We found that by removing the locality condition or local approximation (LA), which was imposed when the imaginary-time evolution operator is converted to a unitary operator, the quantum circuit depth is significantly reduced. We propose two-step approximation methods based on a nonlocality condition: extended LA (eLA) and nonlocal approximation (NLA). To confirm the validity of eLA and NLA, we apply them to the max-cut problem of an unweighted 3-regular graph and a weighted fully connected graph; we comparatively evaluate the performances of LA, eLA, and NLA. The eLA and NLA methods require far fewer circuit depths than LA to maintain the same level of computational accuracy. Further, we developed a “compression” method of the quantum circuit for the imaginary-time steps to further reduce the circuit depth in the QITE method. The eLA, NLA, and compression methods introduced in this study allow us to reduce the circuit depth and the accumulation of error caused by the gate operation significantly and pave the way for implementing the QITE method on NISQ devices.

npj Quantum Information (2021)7:85; <https://doi.org/10.1038/s41534-021-00409-y>

INTRODUCTION

Quantum computers, initially proposed by Feynmann¹, were reported by Benioff², Deutsch^{3,4}, Grover⁵, and Shor⁶ to have great potential that could overwhelmingly surpass that of classical computers. Furthermore, Google experimentally demonstrated quantum supremacy, which is the refutation of the extended Church–Turing thesis, proving the feasibility of quantum computers and raising the expectation for solving practical problems that a classical computer cannot address⁷. Quantum computers can efficiently solve problems in the BQP (bounded-error quantum polynomial) complexity class⁸ and verify an answer to a problem in the QMA (quantum Merlin–Arthur) complexity class⁹. One of the actively researched problems for quantum computers is combinatorial optimization, which is an NP-hard problem¹⁰. Combinatorial optimization problems are closely related to our daily lives, and they include the traveling salesman problem¹¹, scheduling problem¹², and SAT (satisfiability problem) solver¹³. Although combinatorial optimization problems are NP-hard, some quantum algorithms were shown to be superior to the classical ones. Grover’s algorithm is already known to improve the computational cost with quadratic speedup when compared with classical computers^{14,15}. It has been reported that quantum annealing is faster than simulated annealing in several cases^{16–20}. Recently, quantum approximate optimization algorithm (QAOA) has been researched owing to its superiority over classical algorithms, which was demonstrated at the time of its proposal²¹. However, with the development of classical algorithms²², the quantum advantage of QAOA is now an open question.

Under these circumstances, it is challenging for researchers all over the world to employ existing or near-future quantum computers to achieve tasks that are very difficult or impossible using classical computers. Currently available quantum computers are noisy intermediate-scale quantum (NISQ) devices²³. Further,

conventional quantum algorithms, such as Grover’s algorithm, require many gate operations and they cannot be implemented on NISQ devices with no error correction due to short coherence time. Recently, classical-quantum hybrid algorithms called variational quantum eigensolver (VQE)^{24,25}, and QAOA^{21,26–30} have been proposed for NISQ devices. In these methods, ansatz states with parameters are implemented on quantum circuits, and the parameters included in the ansatz states are optimized on a classical computer. While VQE and QAOA can be realized with a limited number of quantum operations and have good noise tolerance, it is difficult to determine the ansatz states properly and converge high-dimensional parameters³¹.

For quantum many-body problems, an imaginary-time evolution method is a known computational method to identify the ground state. The imaginary-time evolution method selectively extracts the ground-state component by performing time evolution in the direction of imaginary time. Various combinatorial optimization problems are converted to a Hamiltonian format, and their corresponding Hamiltonian is derived³². Thus, it is possible to solve the combinatorial optimization problem using the imaginary-time evolution method.

The implementation of the imaginary-time evolution method on a quantum computer involves a critical problem in that the imaginary-time evolution operator is a nonunitary operator, and therefore, it cannot implement the imaginary-time evolution method on a quantum computer in its current state. To overcome this challenge, two quantum imaginary-time evolution (QITE) methods—one that assumes an ansatz state and another that does not—were proposed in previous studies^{33,34}. The method that assumes the ansatz state traces the imaginary-time evolution of the parameters contained in the ansatz state^{33,35,36}. The other method introduces a unitary operation to reproduce the state on

¹Laboratory for Materials and Structures, Institute of Innovative Research, Tokyo Institute of Technology, Tokyo, Japan. ²Quemix Inc., Tokyo, Japan. [✉]email: nishi.h.ac@m.titech.ac.jp

which the imaginary-time evolution operator has acted accurately without assuming an ansatz state^{34,37,38}.

We focus on the QITE method without the ansatz assumption and apply it to the optimization problems. The QITE method requires defining a domain size, which determines the accuracy of reproducing the imaginary-time evolution operator. The quantum circuit of an imaginary-time step scales exponentially with respect to this domain size. Besides, as an additional quantum circuit is added at each imaginary-time step, the quantum circuit becomes deeper in proportion to the lapse of imaginary time³⁴. These two features make it difficult to implement the QITE method on NISQ devices.

Therefore, we propose two approximations and one computational technique to overcome this difficulty. We succeeded in significantly reducing the quantum circuit depth of the QITE method, and we applied the developed algorithms to the max-cut problem, which is an NP-hard problem. For the max-cut problem, we chose an unweighted 3-regular graph and a weighted fully connected graph. The latter is a problem known as the classification problem in the context of unsupervised machine learning^{39,40}.

RESULTS

Unitarization of imaginary-time evolution operators

Consider a scenario wherein a Hamiltonian \hat{H} is given for the optimization problem considered in this study. The Hamiltonian \hat{H} is expressed as the summation of some partial Hamiltonians $\hat{h}[m]$ as $\hat{H} = \sum_{m=1}^{N_{\text{ham}}} \hat{h}[m]$, where N_{ham} is the number of the partial Hamiltonians. The max-cut problem, which is a computational target of this work, is represented by the Hamiltonian in the form of Ising spins and can be mapped to the Pauli-operator representation for qubits in a straightforward manner. In the case

of the Hamiltonian of quantum chemistry, each partial Hamiltonian can be mapped to the Pauli-operator representation on qubits via the Bravyi–Kitaev representation⁴¹ or Jordan–Wigner representation⁴².

For a given Hamiltonian, the ground state is obtained by using the imaginary-time evolution method. We apply the imaginary-time evolution operator defined by $e^{-\tau\hat{H}}$, where τ is the imaginary time to reach the initial ($\tau = 0$) state of the system, $|\Psi(\tau = 0)\rangle$; and $e^{-\tau\hat{H}}|\Psi(\tau = 0)\rangle$. The imaginary-time evolution operator is decomposed by a first-order Suzuki–Trotter decomposition into ones with a small imaginary-time step $\Delta\tau$ ($\tau \equiv \Delta\tau \times N_{\text{step}}$) of the individual partial Hamiltonians $\hat{h}[m]$.

$$e^{-\tau\hat{H}} = \prod_{n=1}^{N_{\text{step}}} \prod_{m=1}^{N_{\text{ham}}} e^{-\Delta\tau\hat{h}[m]} + \mathcal{O}(\Delta\tau^2) . \quad (1)$$

Because the operators of the imaginary-time evolution are nonunitary, they cannot be directly implemented as a gate operation on a quantum computer. In the QITE method, the unitary operator $e^{-i\Delta\tau\hat{A}_n[m]}$ is defined such that it reproduces the state $e^{-\Delta\tau\hat{H}}|\Psi_n\rangle$ for a given state $|\Psi_n\rangle \equiv |\Psi(\tau = n\Delta\tau)\rangle$. We determine the Hermitian operator $\hat{A}_n[m]$ that minimizes the following residual norm.

$$\left\| \frac{e^{-\Delta\tau\hat{h}[m]}|\Psi_n\rangle}{\sqrt{\langle\Psi_n|e^{-2\Delta\tau\hat{h}[m]}|\Psi_n\rangle}} - e^{-i\Delta\tau\hat{A}_n[m]}|\Psi_n\rangle \right\|^2 . \quad (2)$$

Nonlocal condition for imaginary-time evolution operators

We express the Hermitian operator $\hat{A}_n[m]$ as a linear combination of the D th order tensor products of Pauli operators

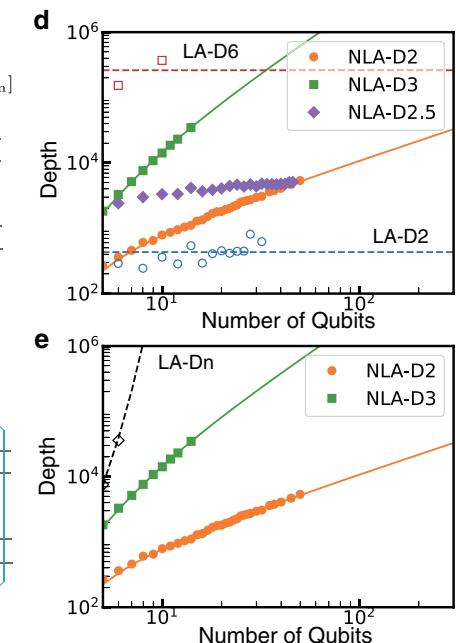
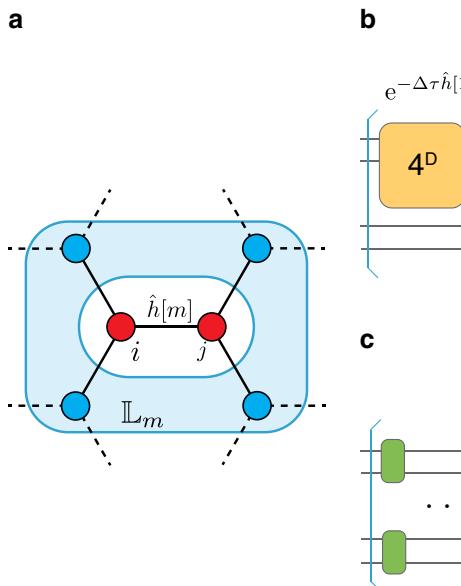


Fig. 1 Quantum circuit diagrams and quantum circuit depth. **a** Schematic of a 3-regular graph. A partial Hamiltonian $\hat{h}[m]$ acts on qubits represented with red vertices, i.e., $\hat{h}[m] \propto \hat{\sigma}_{Zi} \otimes \hat{\sigma}_{Zj}$. Outer blue vertices represent a domain directly connected to the red vertices. Blue vertices contained in the pale blue region comprise the set \mathbb{L}_m in Eq. (3). **b** Quantum circuit diagram for one imaginary-time step of LA. The horizontal line represents each qubit, and the yellow box represents 4^D gating operations on the straddling qubits. **c** Quantum circuit diagram for one imaginary-time step of the NLA (domain size $D = 2$). The green boxes and vertical lines connecting them represent a second-order tensor product operation on the two straddling qubits, with one imaginary-time step containing $N_{\text{bit}} C_D$ of second-order tensor products. Detailed quantum circuit of the two-qubit unitary gate is described in Supplementary Note 2. The dependence of the quantum circuit depth for one imaginary-time step of the max-cut problem in the 3-regular graph (**d**) and the fully connected graph (**e**) as a function of the number of qubits.

Table 1. Scaling of the size of the matrix $S^{(n)}$ and the number of gate operations per qubit of the linear equation of LA and NLA per imaginary-time step.

Method	Scaling of the size of the linear equation	Scaling of the gate operations per qubit
LA	4^D	$4^D N_{\text{ham}} D / N_{\text{bit}}$
NLA	$4^D N_{\text{bit}} C_D$	$4^D N_{\text{bit}} C_{D-1}$

$\{\hat{I}_l, \hat{\sigma}_{X,l}, \hat{\sigma}_{Y,l}, \hat{\sigma}_{Z,l}\}$ acting on the l th qubit as

$$\hat{A}_n[m] = \sum_{l_{k+1}, \dots, l_D \in \mathbb{L}_m} \sum'_{i_1 \dots i_D} a_{i_1 \dots i_D, l_1 \dots l_D}^{(n)}[m] \hat{\sigma}_{i_1, l_1(m)} \otimes \dots \otimes \hat{\sigma}_{i_D, l_D}, \quad (3)$$

where the prime on the first summation symbol indicates removing the double counting of the repeated tensors. We defined \mathbb{L}_m as the set of $N_{\mathbb{L}_m}$ qubits, each of which is directly connected with those acted on by the partial Hamiltonian $\hat{h}[m]$; however, \mathbb{L}_m does not contain the qubits acted on by $\hat{h}[m]$ [see Fig. 1a]. The parameter D , which is called the domain size, satisfies $k D k + N_{\mathbb{L}_m}$, where we assumed the partial Hamiltonian $\hat{h}[m]$ to be written by a tensor product of the k th order. $\{l_1(m), \dots, l_k(m)\}$ is the set of qubits contained in the partial Hamiltonian $\hat{h}[m]$. The summation in Eq. (3) is taken over all combinations of $D - k$ qubits, $\{l_{k+1}(m), \dots, l_D(m)\}$, and chosen from \mathbb{L}_m . D is an input parameter that represents the level of approximation; a larger D indicates that the imaginary-time evolution operator is expressed using higher-order tensor products and the residual norm in Eq. (2) shows a smaller value, which leads to a better approximation. Note that for $D = N_{\text{bit}}$, with N_{bit} being the number of qubits, the residual norm in Eq. (2) vanishes when minimized, yielding the exact imaginary-time evolution operator. In this context, the parameter D represents a truncation level. We consider a scenario where the domain size D incorporates all elements in \mathbb{L}_m , namely $D = k + N_{\mathbb{L}_m}$, and then Eq. (3) reproduces the operator $A_n[m]$ introduced in reference³⁴. This implies that Eq. (3) is a natural extension of the approximation introduced in reference³⁴. We call the method for determining the operator $A_n[m]$ defined in reference³⁴ local approximation (LA) for comparison with later approximation. Then, we refer to the method defined in Eq. (3) as extended local-approximation (eLA). The following notation is used to indicate the domain size D : e.g., LA with $D = 6$ is denoted by LA-D6. Note that, for LA, it is a well-defined approximation only when the domain size $D = k + N_{\mathbb{L}_m}$, and the value of D that can be taken is limited by the Hamiltonian. With an ill-defined domain size D in LA, we found that the calculation accuracy decreased, which is called "Inexact QITE" in reference³⁴. Note that eLA can remove such constraints on the Hamiltonian and flexibly determine the parameter D by considering the linear combination for qubits. This flexibility is obvious in the max-cut problem of the fully connected graph. Solving the minimization problem in Eq. (2) to determine the coefficients $a_{\{i,l\}}^{(n)}[m]$ results in the linear equation $S^{(n)} a^{(n)}[m] = b^{(n)}[m]$, which can be solved using a classical computer. Here, $S_{\{i,l\} \{j,l\}}^{(n)} = \langle \Psi_n | \hat{\sigma}_{\{i,l\}}^\dagger \hat{\sigma}_{\{j,l\}} | \Psi_n \rangle$ and $b_{\{i,l\}}^{(n)}[m] = \langle \Psi_n | \hat{\sigma}_{\{i,l\}}^\dagger \hat{h}[m] | \Psi_n \rangle$. Figure 1b shows a schematic of the quantum circuit representing one imaginary-time step of LA. In LA, the operator of the imaginary-time evolution is approximated by the tensor products of Pauli operators up to the D th order; therefore, 4^D gate operations are required for each partial Hamiltonian. The total number of gate operations for one step of the imaginary-time evolution is $N_{\text{ham}} 4^D$. Table 1 summarizes the size of the linear equation of the LA per step of the imaginary-time evolution and

the number of gate operations per qubit, where N_{bit} is the total number of qubits.

Furthermore, this study proposes another approximation method for \hat{A}_n in the following form:

$$\hat{A}_n = \sum_{l_1, \dots, l_D} \sum_{i_1, \dots, i_D} a_{i_1 \dots i_D, l_1 \dots l_D}^{(n)} \hat{\sigma}_{i_1, l_1} \otimes \dots \otimes \hat{\sigma}_{i_D, l_D}. \quad (4)$$

The difference from Eq. (3) is that we remove the restriction on the set $\{l_1(m), \dots, l_k(m)\}$ and extend the summation over qubits to incorporate all possible combinations of D qubits $\{l_1(m), \dots, l_D(m)\}$. We call this an NLA. As per this definition, we expand the Hermitian operator, \hat{A}_n , using tensor products of Pauli operators over all qubit combinations. Moreover, in LA and eLA, the tensor product space describing $\hat{A}_n[m]$ is different depending on m , which is the partial Hamiltonian. The NLA has a notable feature in that the tensor product space that describes $\hat{A}[m]$ is the same for all m . Table 1 lists the size of the linear equations of the NLA per step of the imaginary-time evolution and the number of gate operations per qubit, where the NLA requires only 4^D unitary operators in $N_{\text{bit}} C_D$ combinations for the quantum circuit in the first step of the imaginary-time evolution. Figure 1c shows the schematic of the quantum circuit of the NLA for one step of the imaginary-time evolution (for $D = 2$).

Reduction effect of circuit depth

To clarify the accuracy and effectiveness of NLA, we applied it to the max-cut problem, which is an NP-hard problem. The Hamiltonian of the max-cut problem in qubit representation is given in the following form containing second-order tensor products³².

$$\hat{H} = - \sum_{(i,j) \in E} d_{ij} \frac{1 - \hat{\sigma}_{Z,i} \hat{\sigma}_{Z,j}}{2} \quad (5)$$

As for the max-cut problem, we considered typical graphs such as 3-regular and fully connected graphs. The 3-regular graphs have three connected edges at every vertex, where E is the set of edges contained in the graph and d_{ij} is the weight of the edges connecting the i th and j th vertices.

The circuit depths, when LA and NLA are applied to the max-cut problem, are shown in Fig. 1d for the 3-regular graph and Fig. 1e for the fully connected graph because different graphs of the max-cut problem change the number of the partial Hamiltonian N_{ham} ; the necessary circuit depths for each approximation change correspondingly. In Fig. 1d, e, the circuit depth calculated using Qiskit⁴³ is plotted with points, and the plotted points are extrapolated. In the case of k -regular graphs, the number of the partial Hamiltonians is given by $N_{\text{ham}} = k N_{\text{bit}}/2$. It increases linearly with the number of vertices N_{bit} so that the number of gate operations per qubit does not depend on the number of qubits, as listed in Table 1. Thus, we extrapolated using $y = \text{const.}$. In NLA, regardless of the structure of the Hamiltonian, the number of gate operations per qubit is scaled by $\mathcal{O}(N_{\text{bit}}^{D-1})$ with respect to the number of qubits N_{bit} because all combinations of $N_{\text{bit}} C_D$ are taken for gate operations including the D th order tensor product. In Fig. 1d, the circuit depth of the NLA is extrapolated by the function fitted by $f(x) = x^{D-1}$.

Note that in LA, $D = 3, 4$, and 5 are not well-defined in the 3-regular graph. Thus, $D = 6$ is required, and $4^6 = 4096$ gate operations are necessary for the imaginary-time evolution of one partial Hamiltonian, which leads to a deeper circuit depth and difficulty in implementation on NISQ devices. In addition, the circuit depth required for LA-D6, compared to NLA-D2, NLA-D3, etc., is considerably higher in the region with a small number of qubits. The circuit depth of the NLA becomes deeper than that of LA in the region where the number of qubits increases.

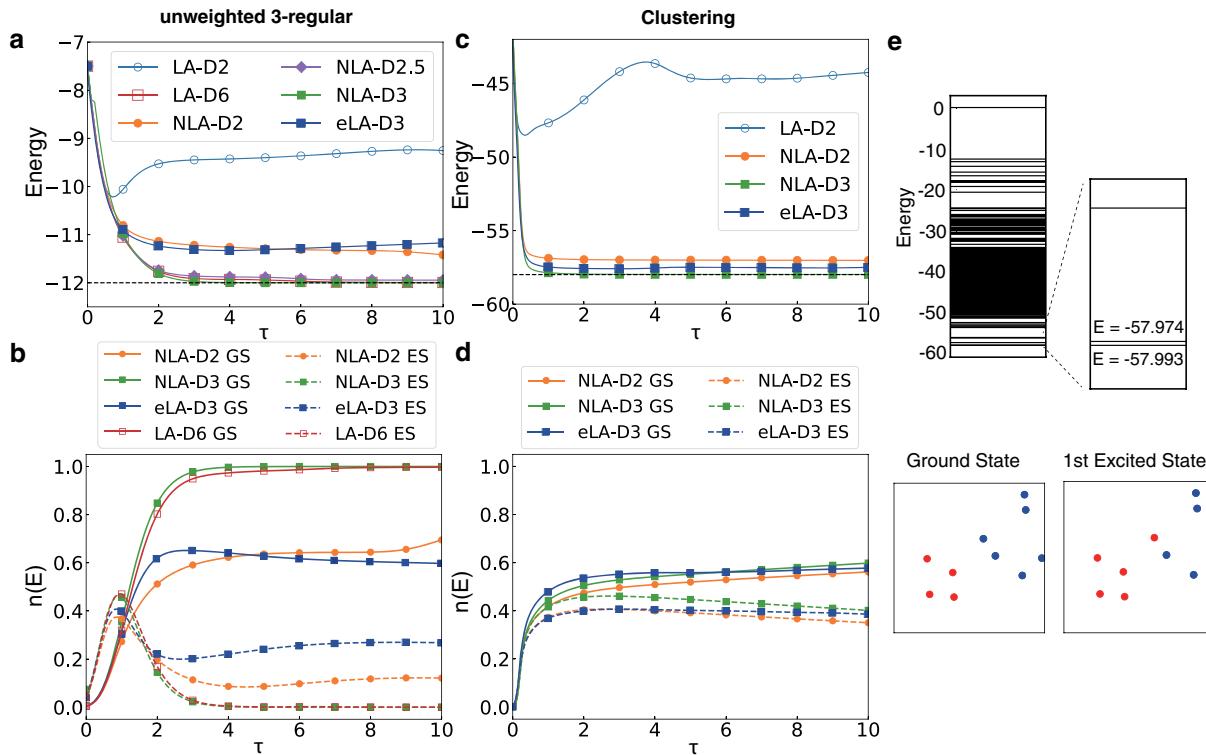


Fig. 2 Numerical simulation of QITE method. Energy E (a) and component proportions of the state $n(E)$ (b) in the QITE method to the max-cut problem for an unweighted 3-regular graph with ten vertices. The ground state is denoted by GS and the first excited state by ES. The energy E (c) and component proportions $n(E)$ (d) of the QITE method for a weighted fully connected graph with ten vertices. e The total energy level diagram of the weighted fully connected graph and the eigenstates corresponding to the ground state and the first excited state (divided into two regions, red and blue).

In Fig. 1e, LA-D2 and eLA-D3 are not shown for the fully connected graph ($N_{\text{ham}}=N_{\text{bit}}C_2$) because the circuit depth of LA-D2 is equal to that of the NLA-D2, and that of eLA-D3 is equal to that of NLA-D3. In addition, because the domain size has to be $D=N_{\text{bit}}$ in LA, which is the exact imaginary-time evolution in a fully connected graph, and the circuit depth increases exponentially with respect to the number of qubits. In NLA, it can be scaled down to the linear or quadratic function with respect to the number of qubits. This result indicates that the NLA and eLA are efficient in reducing the circuit depth, especially when the number of partial Hamiltonians increases; further, these algorithms are effective for NISQ devices.

Calculation accuracy

Simulations were performed after modifying the code provided in reference³⁴. As an initial state, we adopted a state in which all states were superimposed with equal a priori weights. We adopt a figure of merit to discuss the accuracy of the QITE method.

$$r = \lim_{\tau \rightarrow \infty} \frac{\langle \Psi(\tau) | \hat{H} | \Psi(\tau) \rangle}{E_{\text{GS}}}. \quad (6)$$

The first target of the max-cut problem is an unweighted 3-regular graph with ten vertices, where E_{GS} is the energy of the ground state, and it is obtained from the exact diagonalization. The energy of the ground state is $E_{\text{GS}} = -12$. It is known that designing a classical algorithm that achieves $r > 331/332$ for an unweighted 3-regular graph is an NP-hard problem⁴⁴. Further, the approximation accuracy of the current classical algorithm is $r \approx 0.9326$ ⁴⁵. Figure 2a shows the imaginary-time dependence of the energy. The imaginary-time step was set to $\Delta\tau = 0.01$. In LA-D2, as the imaginary-time τ increased, the energy decreased exponentially in the beginning and converged to around -9 , which is

higher than the exact solution by about 3. Another important point is that the energy does not monotonically decreases along the imaginary-time evolution. This behavior indicates that the conversion of the operator of the imaginary-time evolution to the unitary operators is less accurate in expanding it in the space of LA-D2. Furthermore, the LA-D6 calculation result shows $E = -11.99$, which is the energy almost equal to the exact solution. We found that an approximation accuracy in the eLA-D3 is $E = -11.17$ ($r = 0.93$) (the lowest value is $E = -11.33$ ($r = 0.94$)); in NLA-D2, $E = -11.42$ ($r = 0.95$); and in NLA-D3, $E = -12.00$ ($r = 1.00$). We found that eLA-D3 had an approximation accuracy similar to that of the classical algorithm, and NLA-D2 had already exceeded the approximation accuracy of the classical algorithm. NLA-D3 shows better accuracy than NLA-D2 and can reach a nearly exact solution. Note that eLA and NLA monotonically decrease the energy along the imaginary-time evolution with sufficiently good accuracy compared to LA-D2. This behavior was confirmed not only for NLA-D2 but also for NLA-D3 and others. As can be seen from Fig. 1d, in LA-D6, the circuit depth of one imaginary-time step is 369757, while the circuit depth in the NLA-D2 is 789. This implies the circuit depth of NLA can be significantly shallower than that of LA.

While NLA-D3 has extremely high accuracy, its circuit depth increases with a quadratic function with respect to the number of qubits. Then, we developed NLA-D2.5 to keep the scaling of the circuit depth as linear as NLA-D2 while maintaining the accuracy of NLA-D3, which is an approximation to expand the space of \hat{A}_n to the space involving the second-order tensor products incorporated by NLA-D2 and the third-order tensor products by eLA-D3. Thus, by incorporating some portions of bases of eLA-D3 into those of NLA-D2, computational scaling can be made linear with respect to the number of qubits, which makes it applicable even in regions with a large number of qubits. Figure 1d shows

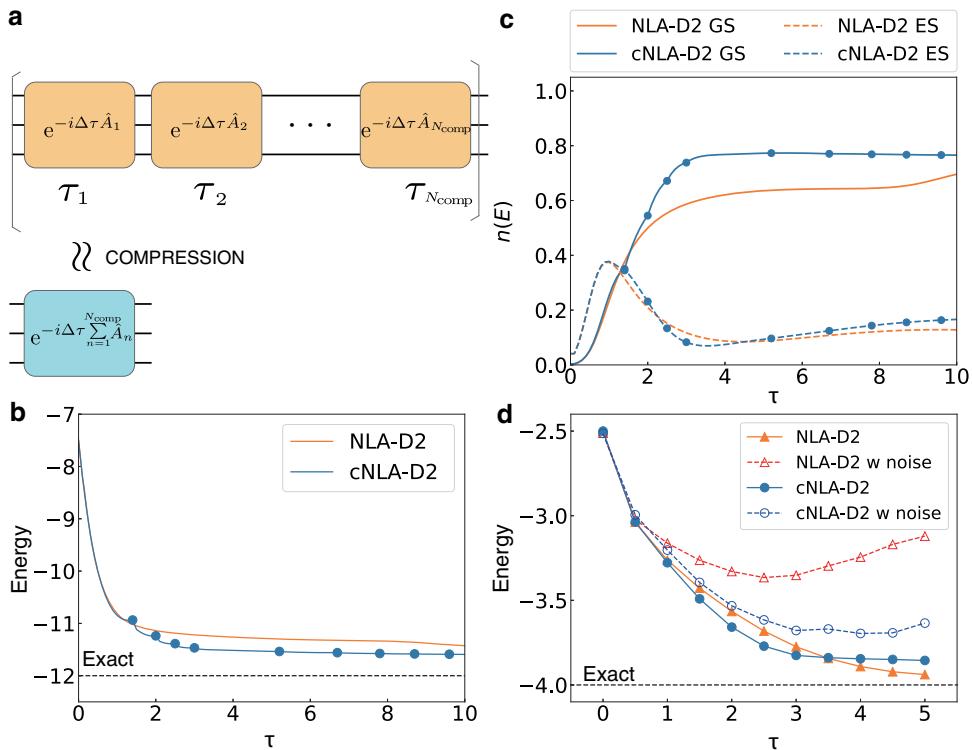


Fig. 3 Numerical simulation of compression of imaginary-time steps. **a** Schematic of the compression of the imaginary-time step. N_{comp} steps are compressed into one step by 1st-order Suzuki-Trotter decomposition. Energy E (**b**) and component of the eigenstate $n(E)$ (**c**) in the imaginary-time evolution with and without compression of the imaginary-time step in the max-cut problem of an unweighted 3-regular graph with ten vertices. The compressed point N_{comp} is plotted with circles. **d** Results of the simulation with noise for the max-cut problem in an unweighted graph with four vertices.

that the circuit depth is almost the same as that of NLA-D2 for 50 qubits or more, which means that the circuit depth can be significantly reduced compared to that of NLA-D3. In addition, the calculation result of NLA-D2.5 is $E = -11.95$ and $r = 0.99$, which gives a good approximation accuracy with a small circuit depth.

Here, for further consideration, we decomposed the state $|\Psi(\tau)\rangle = e^{-i\hat{H}\tau}|\Psi(\tau=0)\rangle$ into the eigenstate components of the Hamiltonian, and the calculated $n(E) \equiv \sum_i |\langle i|\Psi(\tau)\rangle|^2 \delta(E - E_i)$ as a function of energy E at each imaginary-time step τ is plotted in Fig. 2b where $|i\rangle$ is the eigenstate of \hat{H} and E_i is the eigen energy of $|i\rangle$. Here, we note that the ground state of the eigenstate component $n(E_{\text{GS}})$ is equal to the so-called fidelity defined as $F = |\langle \Psi(\tau) | \Psi_{\text{GS}} \rangle|^2$. The ground state can be observed with probabilities of $n(E_{\text{GS}}) = 0.60$ for eLA-D3 (at maximum, $n(E_{\text{GS}}, \tau=2.87) = 0.65$), $n(E_{\text{GS}}) = 0.69$ for NLA-D2, $n(E_{\text{GS}}) = 0.97$ for NLA-D2.5, and $n(E_{\text{GS}}) = 1.00$ for NLA-D3. The imaginary-time dependence of the probability of the first excited state is also plotted. For the first excited state, it is observed that the probability is amplified up to $\tau=1$, and it starts to decrease, which increases the ground-state probability.

Next, we deal with another computational model called a weighted fully connected graph (classification problem). The coupling constants d_{ij} were given by random numbers. The ground-state energy is $E_{\text{GS}} = -57.993$. In addition, the imaginary-time step is set to $\Delta\tau = 0.01$. In the classification problem, as shown in Fig. 2e, each graph vertex is colored red or blue. In LA-D2, as in the 3-regular graph, we observed that the energy does not necessarily decrease monotonically. The energy of eLA-D3 is lower than that of NLA-D2; $E = -57.504$ ($r = 0.99$) for eLA-D3, $E = -57.026$ ($r = 0.98$) for NLA-D2, and $E = -57.985$ ($r = 0.99$) for NLA-D3 (Fig. 2c). From the viewpoint of the component analyses of the states, the ground state and the first excited state are pseudo-degenerate (Fig. 2e), and therefore, the probability of the first

excited state remains at the same level as the ground state even around $\tau = 2$ when the energy converges sufficiently (Fig. 2d). In NLA, the first excited state gradually decays along with the imaginary-time evolution; however, a sufficiently long imaginary-time evolution is necessary. In particular, NLA-D2 behaves similarly to NLA-D3, and NLA-D2 is sufficiently accurate to obtain the ground state in actual applications.

We now consider why the accuracy of eLA and NLA is better than that of LA with a relatively small domain size D . From the actual application results of eLA-D3, we found that the $b_{\{i,l\}}^{(n)}[m] = \langle \Psi_n | \hat{\sigma}_{\{i,l\}}^\dagger \hat{h}[m] | \Psi_n \rangle \approx 0$ when the Pauli operator $\hat{\sigma}_{\{i,l\}}^\dagger$ and $\hat{h}[m]$ do not intersect each other. With a rough approximation for such cases, $b_{\{i,l\}}^{(n)}[m] = \langle \Psi_n | \hat{\sigma}_{\{i,l\}}^\dagger \hat{h}[m] | \Psi_n \rangle = 0$, a sparsity in the coefficients of \hat{A}_n can be deduced, which eLA highlight and leverage. This fact means that the terms that would require a large domain size D in LA can be efficiently captured with a smaller domain size D in eLA, leading to its high accuracy. Furthermore, by considering that the definition of NLA is expanded from that of eLA, NLA can improve further the accuracy of eLA.

Compression of imaginary-time steps

The approximation accuracy of the NLA and its circuit depth have been discussed. The ‘compression of imaginary-time steps’ is introduced in this section for further reduction of the number of gate operations in NLA. Figure 3a shows a schematic of the compression technique. When the imaginary-time step $\Delta\tau$ is sufficiently small, the time-evolution operators can be compressed into a single exponential form via the reverse Suzuki-Trotter decomposition

$$\prod_{n=1}^{N_{\text{comp}}} \exp(-i\Delta\tau\hat{A}_n) = \exp(-i\Delta\tau \sum_{n=1}^{N_{\text{comp}}} \hat{A}_n) + \mathcal{O}(\Delta\tau^2), \quad (7)$$

where N_{comp} is the number of compressed steps. It is necessary to choose an appropriate N_{comp} within the range that guarantees sufficient accuracy for the Suzuki–Trotter decomposition because its accuracy decreases if the N_{comp} becomes large. To determine the specific N_{comp} in this work, we increased the N_{comp} parameter by one at every time-evolution step until the total energy increases. In actual QITE calculations, N_{comp} is not necessarily a constant throughout the calculation. This method enables the reduction of quantum circuits to as small as $1/N_{\text{comp}}$. We discussed the error of the second order for $\Delta\tau$ in the compression method in Supplementary Note 1.

The graph used for the calculation is the same as that in Fig. 2a, b, which is a 3-regular graph with ten vertices. Figure 3 shows the results of the compression technique for the QITE. In Fig. 3b, the time the compression ended is plotted as a blue circle. In the case of Fig. 3b, the quantum circuit depth is significantly reduced by the compression technique to four compressed imaginary-time steps, and the energy at $\tau=10$ is $E=-11.43$ ($r=0.95$) without and $E=-11.59$ ($r=0.97$) with the compression technique. We found that sufficient accuracy was achieved regardless of the compression, which indicates that compression does not affect the results. It may be assumed that the compressed technique has a lower energy than that of the uncompressed calculation; a detailed investigation revealed that this was attributed to the accidental acceleration of the convergence by compression. Figure 3c plots the component analyses of the wavefunctions during the imaginary-time evolution with and without the compression method. Finally, the probability of obtaining a ground state is $n(E_{\text{min}})=0.76$ with and $n(E_{\text{min}})=0.73$ without the compression technique.

The “compression of imaginary-time steps” is effective in reducing the circuit depth, and simultaneously, it reduces the noise associated with the gate operations. We discuss the results of the simulation with noise. The actual qubits are currently connected only with neighboring sites; however, in this study, we simulated a fully connected model. For implementation on an actual quantum computer, in which only adjacent sites are connected, a SWAP gate can be used with an overhead of $\mathcal{O}(\sqrt{N_{\text{bit}}})$ ⁴⁶. For example, QAOA uses a SWAP network^{47,48} to implement a $\mathcal{O}(N_{\text{bit}})$ overhead⁴⁹. We describe the quantum circuit of NLA-D2 of the QITE method for an adjacent-coupling circuit using the SWAP network in Supplementary Note 2. Figure 3d shows the simulation results of the max-cut problem for an unweighted graph with four vertices. The coefficients $a_{\{i,j\}}^{(n)}$ in Eq. (4) for the noisy calculation are the same as those for the non-noisy calculation. The noiseless condition without compression results in $E=-3.94$, which is close to the exact solution $E=-4.00$ around $\tau=5$. However, the circuit depth is 922 ($\Delta\tau=0.5$), and the simulation result with noise is $E=-3.13$, which is far from the exact solution. This gap was attributed to the accumulation of errors caused by an increase in circuit depth. The result with compression is $E=-3.85$ in the case without noise; however, the circuit depth is 163, and the effect of noise is expected to be less sensitive. In fact, the simulation result with noise is $E=-3.63$, which shows that the noise can be reduced with compression. Thus, it has been shown that the “compression” method of quantum circuits has the advantage of reducing the accumulation of errors.

DISCUSSION

In this study, we proposed two-step approximation methods based on nonlocality: eLA and NLA. We applied them to the Max-cut problem of an unweighted 3-regular graph and a weighted fully connected graph, and comparatively validated the performances of LA, eLA, and NLA. We found that NLA requires significantly less circuit depth than LA while maintaining the same level of computational accuracy. For example, when we request

the classical approximation limit in the QITE calculations, the circuit depth required for a single imaginary-time step can be significantly reduced from 369,757 for LA to 789 for NLA when applying it to a 3-regular graph, and from about 314,000 for LA to 789 for NLA when applying it to a fully connected graph. Further, we developed a “compression” technique of the imaginary-time evolution steps to further reduce the circuit depth in the QITE method. With this compression method, we succeeded in further reducing the circuit depth. We showed that the reduction in circuit depth using this compression method has a secondary effect of reducing the accumulation of error caused by the gate operation. Thus, it is an effective method for realization on NISQ devices. The eLA, NLA, and compression methods introduced in this study enable us to significantly reduce the circuit depth and the accumulation of error caused by the gate operation and have paved the way for the realization of the QITE method on NISQ devices.

METHODS

Noisy simulation of QITE method

Our numerical simulations were performed after implementing the eLA, NLA, and compression method on the code provided in reference³⁴. The simulation of the quantum noise’s presence is performed with the implementation of the QITE method at the level of the NLA-D2 on the IBM Qiskit quantum simulator. Although almost all actual quantum devices’ qubit connectivity is restricted, we simulated the QITE method based on the fully connected coupling. For implementation on a device connected only with neighboring qubits, we provide a circuit of the QITE method using the SWAP network in Supplementary Note 2. The error model of the gate was constructed from the thermal relaxation time ($T_1, T_2=(100\text{ }\mu\text{s}, 80\text{ }\mu\text{s})$, and the gate time ($T_{g1}, T_{g2}=(0.02\text{ ns}, 0.1\text{ ns})$). The noise simulation was performed by introducing the readout errors ($p_{00}, p_{01}, p_{10}, p_{11}=(0.995, 0.005, 0.02, 0.98)$). These parameters were assumed to be close to the actual values of IBMQ⁵⁰.

Note added to proof

During our review of this paper, we noticed an independent work-related “compression method” being done in parallel⁵¹.

DATA AVAILABILITY

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

CODE AVAILABILITY

The code developed for the current study is available from the corresponding author on reasonable request.

Received: 3 June 2020; Accepted: 14 April 2021;

Published online: 01 June 2021

REFERENCES

1. Feynman, R. P. Simulating physics with computers. *Int. J. Theor. Phys.* **21**, 467–488 (1982).
2. Benioff, P. The computer as a physical system: a microscopic quantum mechanical hamiltonian model of computers as represented by turing machines. *J. Stat. Phys.* **22**, 563–591 (1980).
3. Deutsch, D. Quantum theory, the church–turing principle and the universal quantum computer. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **400**, 97–117 (1985).
4. Deutsch, D. & Jozsa, R. Rapid solution of problems by quantum computation. *Proc. R. Soc. Lond. Ser. A: Math. Phys. Sci.* **439**, 553–558 (1992).
5. Grover, L. K. A fast quantum mechanical algorithm for database search. In *Proc. Twenty-eighth Annual ACM Symposium on Theory of Computing*, 212–219 (1996).
6. Shor, P. W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Rev.* **41**, 303–332 (1999).

7. Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).
8. Bernstein, E. & Vazirani, U. Quantum complexity theory. *SIAM J. Comput.* **26**, 1411–1473 (1997).
9. Kitaev, A. Y., Shen, A., Vyalyi, M. N. & Vyalyi, M. N. *Classical and Quantum Computation*. 47 (American Mathematical Soc., 2002).
10. Hyafil, L. & Rivest, R. L. *Graph Partitioning and Constructing Optimal Decision Trees are Polynomial Complete Problems* (IRIA, Laboratoire de Recherche en Informatique et Automatique, 1973).
11. Lawler, E. L. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics (John Wiley & Sons, 1985).
12. Ho, N. B. & Tay, J. C. Genace: an efficient cultural algorithm for solving the flexible job-shop problem. In *Proc. 2004 Congress on Evolutionary Computation (IEEE Cat. No. 04TH8753)*, Vol. 2, 1759–1766 (IEEE, 2004).
13. Karp, R. M. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, 85–103 (Springer, 1972).
14. Durr, C. & Hoyer, P. A quantum algorithm for finding the minimum. arXiv. [quant-ph/9607014](https://arxiv.org/abs/9607014) (1996).
15. Baritompa, W. P., Bulger, D. W. & Wood, G. R. Grover's quantum algorithm applied to global optimization. *SIAM J. Optim.* **15**, 1170–1184 (2005).
16. Kadowaki, T. & Nishimori, H. Quantum annealing in the transverse ising model. *Phys. Rev. E* **58**, 5355 (1998).
17. Farhi, E. et al. A quantum adiabatic evolution algorithm applied to random instances of an np-complete problem. *Science* **292**, 472–475 (2001).
18. Johnson, M. W. et al. Quantum annealing with manufactured spins. *Nature* **473**, 194–198 (2011).
19. Albash, T. & Lidar, D. A. Demonstration of a scaling advantage for a quantum annealer over simulated annealing. *Phys. Rev. X* **8**, 031016 (2018).
20. Susa, Y. et al. Quantum annealing of the p-spin model under inhomogeneous transverse field driving. *Phys. Rev. A* **98**, 042326 (2018).
21. Farhi, E., Goldstone, J. & Gutmann, S. A quantum approximate optimization algorithm. arXiv. <https://arxiv.org/abs/1411.4028> (2014).
22. Barak, B. et al. Beating the random assignment on constraint satisfaction problems of bounded degree. arXiv. <https://arxiv.org/abs/1505.03424> (2015).
23. Preskill, J. Quantum computing in the nisq era and beyond. *Quantum* **2**, 79 (2018).
24. Peruzzo, A. et al. A variational eigenvalue solver on a photonic quantum processor. *Nat. Commun.* **5**, 4213 (2014).
25. McClean, J. R., Romero, J., Babbush, R. & Aspuru-Guzik, A. The theory of variational hybrid quantum-classical algorithms. *N. J. Phys.* **18**, 023023 (2016).
26. Otterbach, J. et al. Unsupervised machine learning on a hybrid quantum computer. arXiv. <https://arxiv.org/abs/1712.05771> (2017).
27. Moll, N. et al. Quantum optimization using variational algorithms on near-term quantum devices. *Quantum Sci. Technol.* **3**, 030503 (2018).
28. Wang, Z., Hadfield, S., Jiang, Z. & Rieffel, E. G. Quantum approximate optimization algorithm for maxcut: a fermionic view. *Phys. Rev. A* **97**, 022304 (2018).
29. Hadfield, S. et al. From the quantum approximate optimization algorithm to a quantum alternating operator ansatz. *Algorithms* **12**, 34 (2019).
30. Guerreschi, G. G. & Matsueda, A. Y. Qaoa for max-cut requires hundreds of qubits for quantum speed-up. *Sci. Rep.* **9**, 6903 (2019).
31. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R. & Neven, H. Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **9**, 1–6 (2018).
32. Lucas, A. Ising formulations of many np problems. arXiv. <https://arxiv.org/abs/1302.5843> (2014).
33. McArdle, S. et al. Variational ansatz-based quantum simulation of imaginary time evolution. *npj Quantum Inf.* **5**, 1–6 (2019).
34. Motta, M. et al. Determining eigenstates and thermal states on a quantum computer using quantum imaginary time evolution. *Nat. Phys.* **16**, 205–210 (2020).
35. Stokes, J., Izrae, J., Killor, N. & Carleo, G. Quantum natural gradient. arXiv. <https://arxiv.org/abs/1909.02108> (2019).
36. David, W., Christian, G. & Michael, K. Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizert. arXiv. <https://arxiv.org/abs/2004.14666> (2020).
37. Yeter-Aydeniz, K., Pooser, R. C. & Siopsis, G. Practical quantum computation of chemical and nuclear energy levels using quantum imaginary time evolution and lanczos algorithms. arXiv. <https://arxiv.org/abs/1912.06226> (2019).
38. Beach, M. J., Melko, R. G., Grover, T. & Hsieh, T. H. Making trotters sprint: a variational imaginary time ansatz for quantum many-body systems. *Phys. Rev. B* **100**, 094434 (2019).
39. Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: a review. *ACM Comput. Surv. (CSUR)* **31**, 264–323 (1999).
40. Jain, A. K. & Dubes, R. C. *Algorithms for Clustering Data* (Prentice-Hall, Inc., 1988).
41. Bravyi, S. B. & Kitaev, A. Y. Fermionic quantum computation. *Ann. Phys.* **298**, 210–226 (2002).
42. Jordan, P. & Wigner, E. P. Über das paulische Äquivalenzverbot. *Eur. Phys. J.* **47**, 631–651 (1928).
43. Aleksandrowicz, G. et al. Qiskit: an open-source framework for quantum computing. Accessed 16 Mar 2019.
44. Berman, P. & Karpinski, M. On some tighter inapproximability results (extended abstract). In *Proc. 26th International Colloquium on Automata, Languages and Programming*, 200–209 (1999).
45. Goemans, M. X. & Williamson, D. P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM* **42**, 1115–1145 (1995).
46. Cheung, D., Maslov, D. & Severini, S. Translation techniques between quantum circuit architectures. In *Proc. Workshop on Quantum Information Processing* (2007).
47. Kivlichan, I. D. et al. Quantum simulation of electronic structure with linear depth and connectivity. *Phys. Rev. Lett.* **120**, 110501 (2018).
48. Babbush, R. et al. Low-depth quantum simulation of materials. *Phys. Rev. X* **8**, 011044 (2018).
49. Crooks, G. E. Performance of the quantum approximate optimization algorithm on the maximum cut problem. arXiv. <https://arxiv.org/abs/1811.08419> (2018).
50. IBM Quantum Experience Web Site. <https://quantum-computing.ibm.com/>.
51. Gomes, N. et al. Efficient step-merged quantum imaginary time evolution algorithm for quantum chemistry. *J. Chem. Theory Comput.* **16**, 6256–6266 (2020).

ACKNOWLEDGEMENTS

This research was supported by MEXT as an Exploratory Challenge on Post-K computer (Frontiers of Basic Science: Challenging the Limits) and by Grants-in-Aid for Scientific Research (A) (Grant Number 18H03770) from JSPS (Japan Society for the Promotion of Science).

AUTHOR CONTRIBUTIONS

H.N. and Y.M. conceived the general idea. H.N. modified the code provided in prior work. H.N. and T.K. developed the code for noisy simulation. Numerical simulations were performed by H.N. All authors contributed equally to the manuscript preparation and presentation of results.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41534-021-00409-y>.

Correspondence and requests for materials should be addressed to H.N.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.