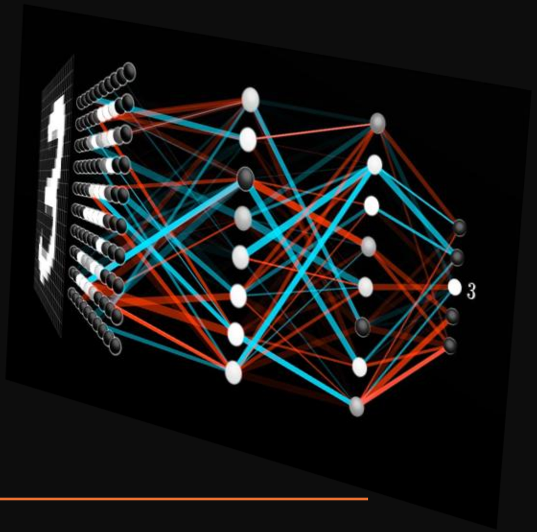
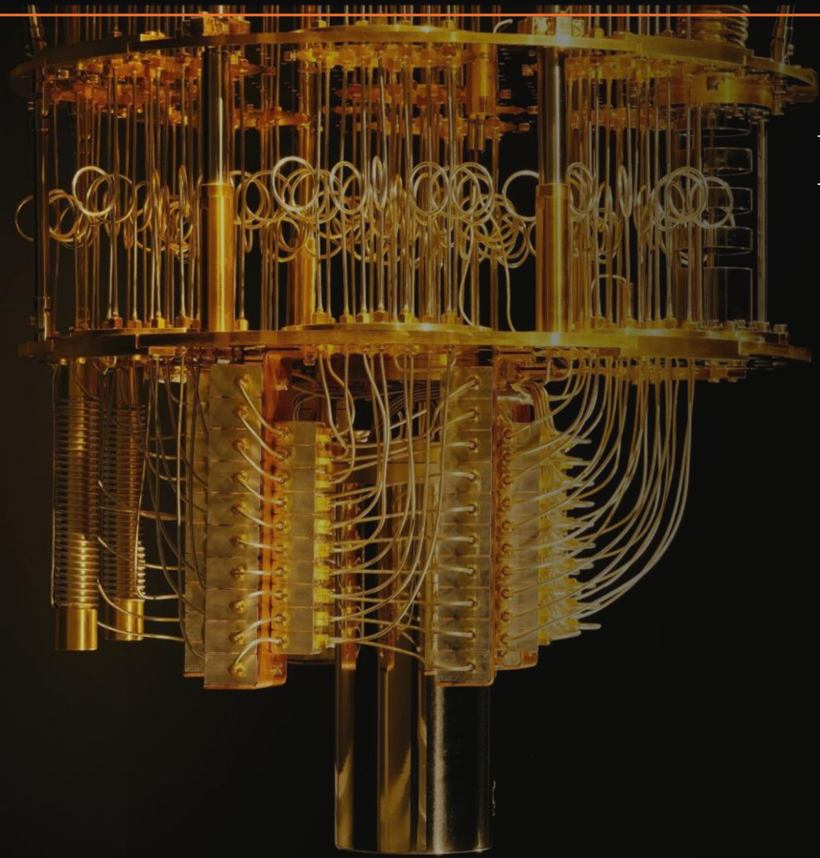


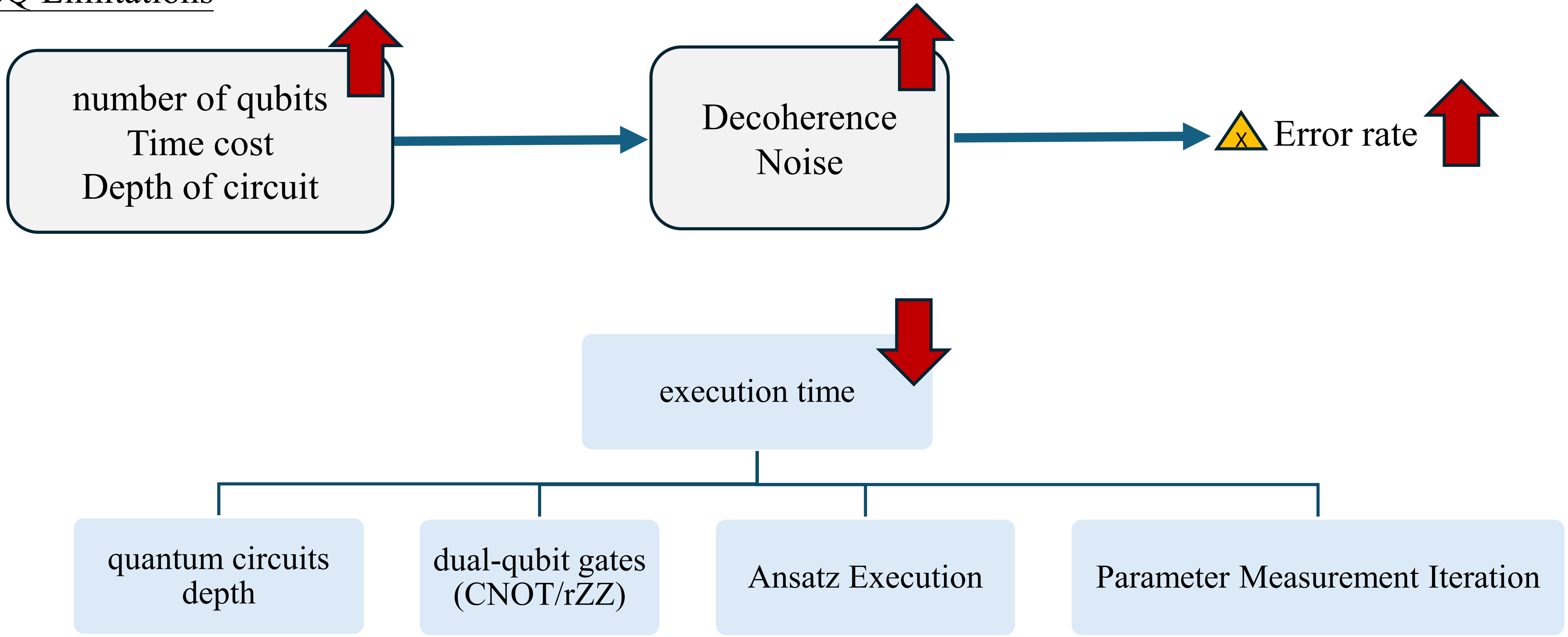


Learning Efficient Variational Quantum Circuits With Deep Reinforcement Learning

Team Number：9

Theme background and motivation

<u>NISQ Limitations</u>



Variational Quantum Eigensolver , VQE

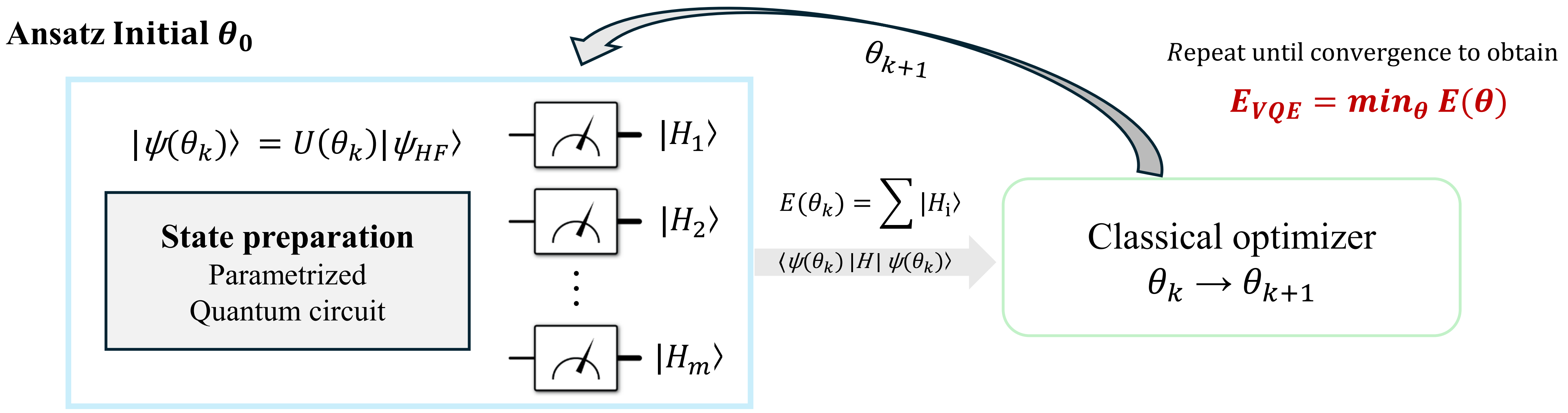
Hybrid Quantum-Classical

Quantum subroutine

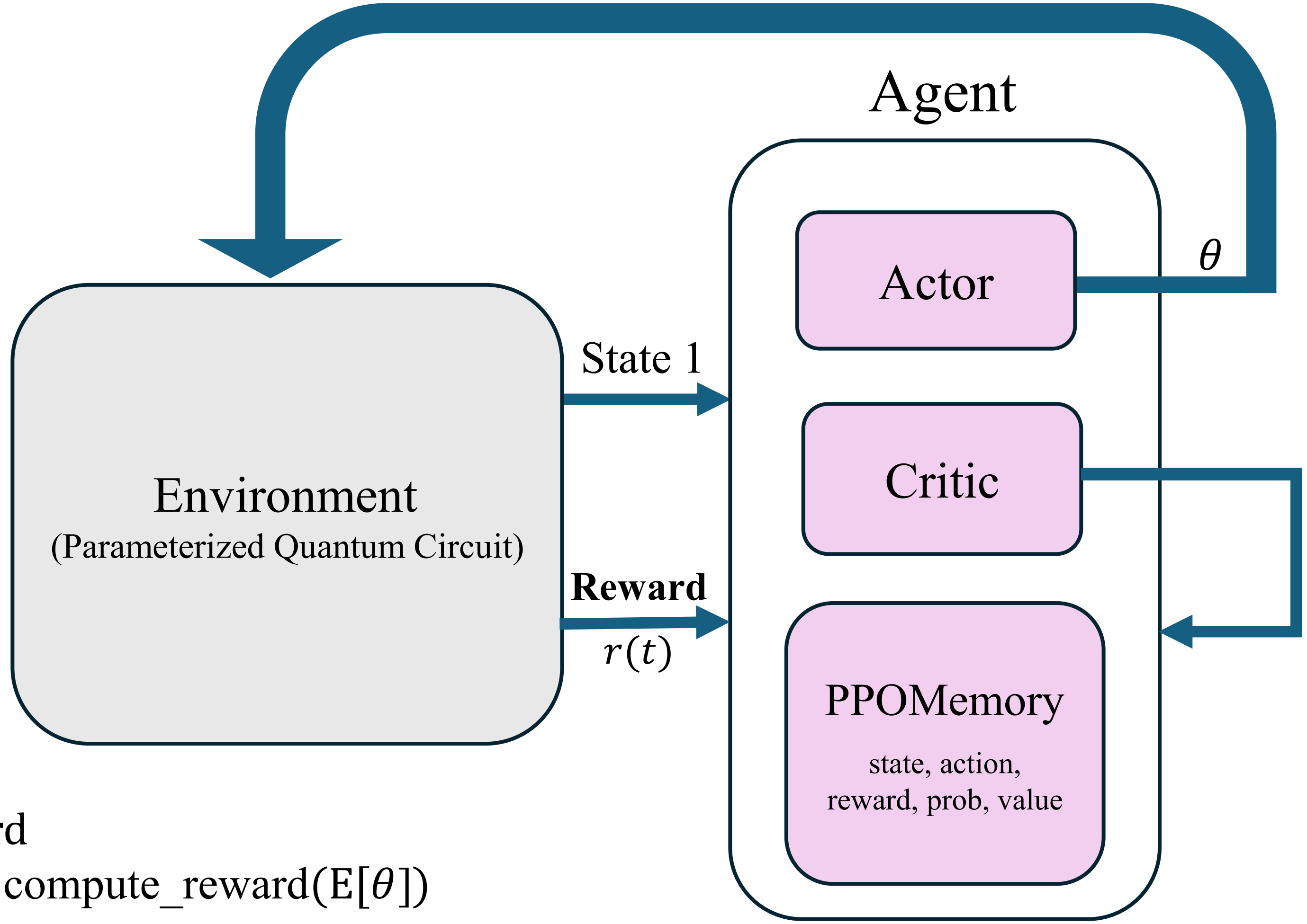
- Parameterized Quantum Circuit (PQC)
- measures the Hamiltonian

Classical optimized

- Updates the circuit parameters iteratively



Proximal Policy Optimization PPO-based optimizer

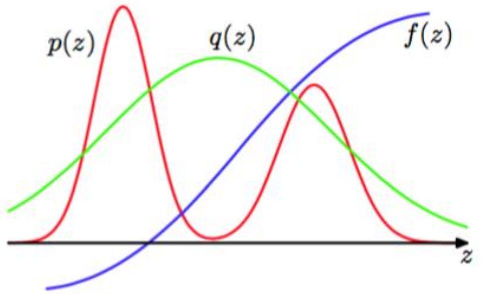


★ $E(\theta) = \langle \psi(\theta) | H | \psi(\theta) \rangle$

$\downarrow$

$\theta^{*} = \arg\min E\,(\theta)$

Expected Return R_θ



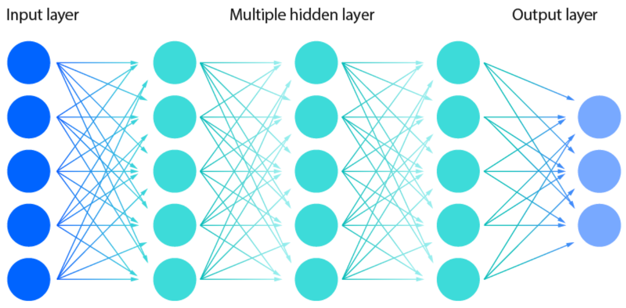
Reward

$r(t)$ = compute_reward(E[θ])

$$Expected\ Return\ R_\theta = \textstyle\sum_\tau R(\tau)\, p_\theta(\tau) = E_{\tau \sim p_\theta}[R(\tau)] \Rightarrow R(\tau) = \sum_{t=1}^{T} r(t)$$

$$E_{x \sim p}\,[f(x)] = E_{x \sim p}\,[f(x)\frac{p(x)}{q(x)}]$$

Project Strategy — VQE with PPO



1. reinforcement learning (DRL): Apply DRL algorithms for quantum circuit design
2.Design efficient variational quantum circuits:
Optimize quantum circuits for better performance and resource efficiency
3.Apply to molecular ground state problems:
Use the optimized circuits to accurately estimate the ground state energy of molecular systems

Quantum subroutine

- Parameterized Quantum Circuit
- Measures the Hamiltonian

Classical optimized

- PPO-based optimizer
- Updates the circuit parameters iteratively

Agent

PQC parameters → Execute PQC → Measur Hamiltonian → Critic evaluates value Compute Advantage → Actor Updates Policy → **The BEST**

Reward from Energy

Renew PQC parameter (based on updated policy)

Molecular and Quantum Chemistry Settings

```
[MOL]
mol_name = LiH
atoms = ["Li", "H"]
coordinates = ([0.0, 0.0, 0.0], [0.0, 0.0, 1.571274961436279])
multiplicity = 1
charge = 0
num_electrons = 4
num_spatial_orbitals = 6
num_particles = (2, 2)
num_qubits = 12
fci_energy = -7.88266974664723
```

DPO Configuration Settings

```
[DPO]
use_dpo = True
dpo_beta = 0.1
dpo_loss_weight = 0.5
reference_update_freq = 5
preference_buffer_size = 1000
```

Reinforcement Learning Training Parameters

```
[TRAIN]
learning_rate = 0.0003
gamma = 0.99
gae_lambda = 0.95
policy_clip = 0.2
batch_size = 64
num_episodes = 1000
num_steps = 20
num_epochs = 10
max_circuit_depth = 50
conv_tol = 1e-5
optimizer_option = "Adam"
```

System Parameters / Experimental Setup

Parameterized Quantum Circuit

$$\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}_{50X17}$$

- gate_type_one_hot(12 gate)
- target_qubit
- control_qubit
- angle
- position
- connectivity_flag

gate_type	gate_type_one_hot
H	100000000000
X	010000000000
Y	001000000000
Z	000100000000
Cx	000010000000
Cz	000001000000
Rx	.
Ry	.
Rz	.
T	
S	
SX	

Gate type

```python
# Supported gate types and their indices
gate_types = ['h', 'x', 'y', 'z', 'cx', 'cz', 'rx', 'ry', 'rz', 't', 's', 'sx']
num_gate_types = len(gate_types)
```

ibm_pittsburgh	us-east	156
ibm_kingston	us-east	156
ibm_fez	us-east	156
ibm_marrakesh	us-east	156
ibm_torino	us-east	133
ibm_brisbane	us-east	127

Items per page: 10 1–6 of 6 items

ibm_pittsburgh

You do not have access to this system in this account. Switch accou

Status
Online

Region
Washington DC (us-east)

Basis gates
cz, id, rx, rz, rzz, sx, x

Total pending jobs
46

Median readout error
4.15E-3

Median T1
317.91 us

Calibration data

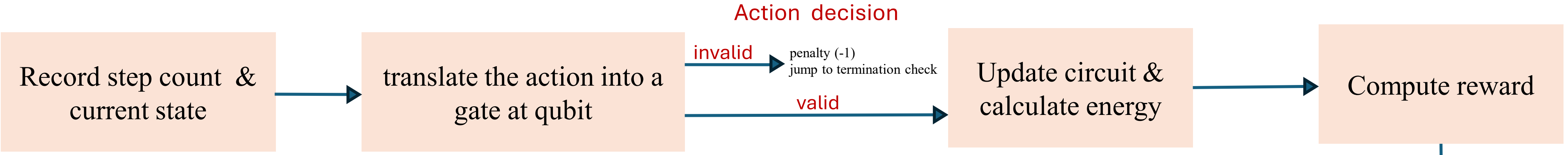
Map view | Graph view | Table view | Expand

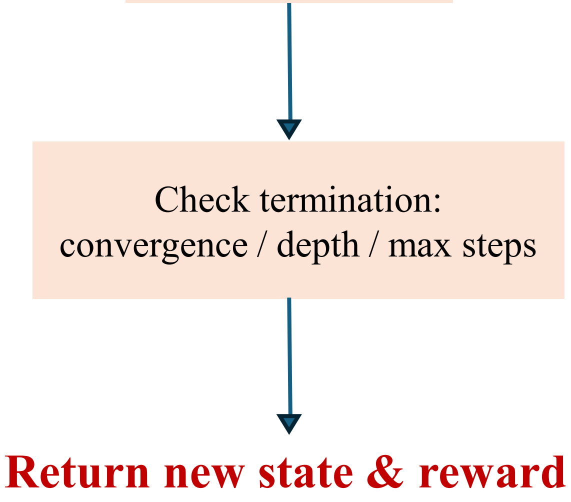
Reward function

- **compute_reward**

$$\Delta E = |E - E_{FCI}|$$

$$reward = accuracy\_reward - complexity\_penalty + convergence\_bonus$$
$$= -100000\,\Delta E - 0.01(\#CNOT + \mathrm{depth} + \#\mathrm{gates}) + 10[\Delta E < conv\_tol]$$





- Encourage accuracy → Get close to target energy
- Discourage waste → Keep the circuit simple
- Reward completion → Extra bonus for early convergence
- Punish invalid actions → Avoid wasted moves

Categorical Activation Function

Purpose

- decision-making processes for **discrete** action spaces
- Suitable for multi-class classification problems.

Principle

The final layer produces **logits** (raw scores).

Logits are converted into a **probability distribution** using the **softmax** function.

Each class/action corresponds to one probability value.

Calculation:

Softmax formula:

$$P(y = i) = \frac{e^{Zi}}{\sum_{j=1}^{K} e^{Zj}}$$

Zi is the logit of the i-th class and K is the total number of classes

Common Application:

In policy networks, outputs the action probability distribution.The agent selects actions based on these probabilities.

State Space & Action Space in RL

State Space The complete set of all possible observations the agent can perceive from the environment.

In this proj：

- Gate type (one-hot encoding)
- Target qubit
- Control qubit
- Rotation angle
- Gate position
- Connectivity flag

Where am I now ?

A state is the **current quantum circuit** represented as an encoded matrix/vector

Action Space The complete set of possible moves the agent can take in each state.

In this proj：

- Discrete actions for modifying the quantum circuit:
    - Select gate type (e.g., Rx, Ry, Rz, H, CNOT, CZ...)
    - Assign target qubit
    - Assign control qubit (if required)
    - Set rotation parameters

Chosen using a **categorical activation function** to model the action distribution

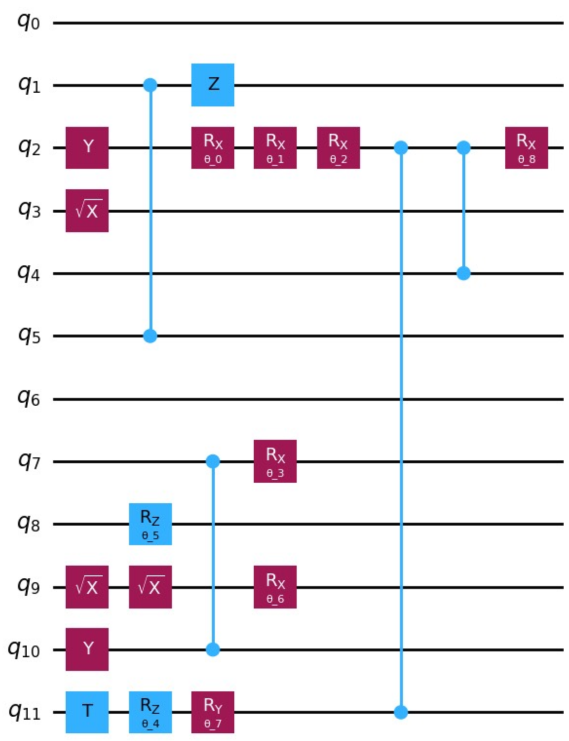
What can I do next ?

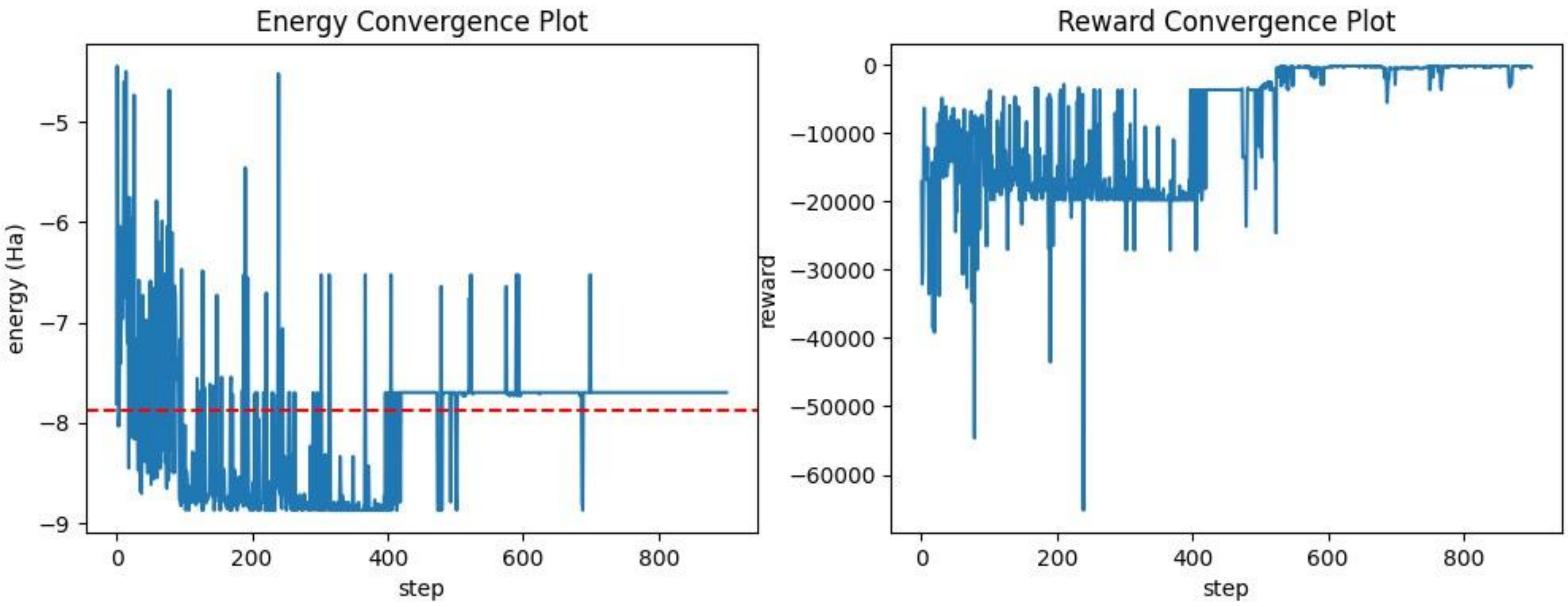
Result

Simulation Environments

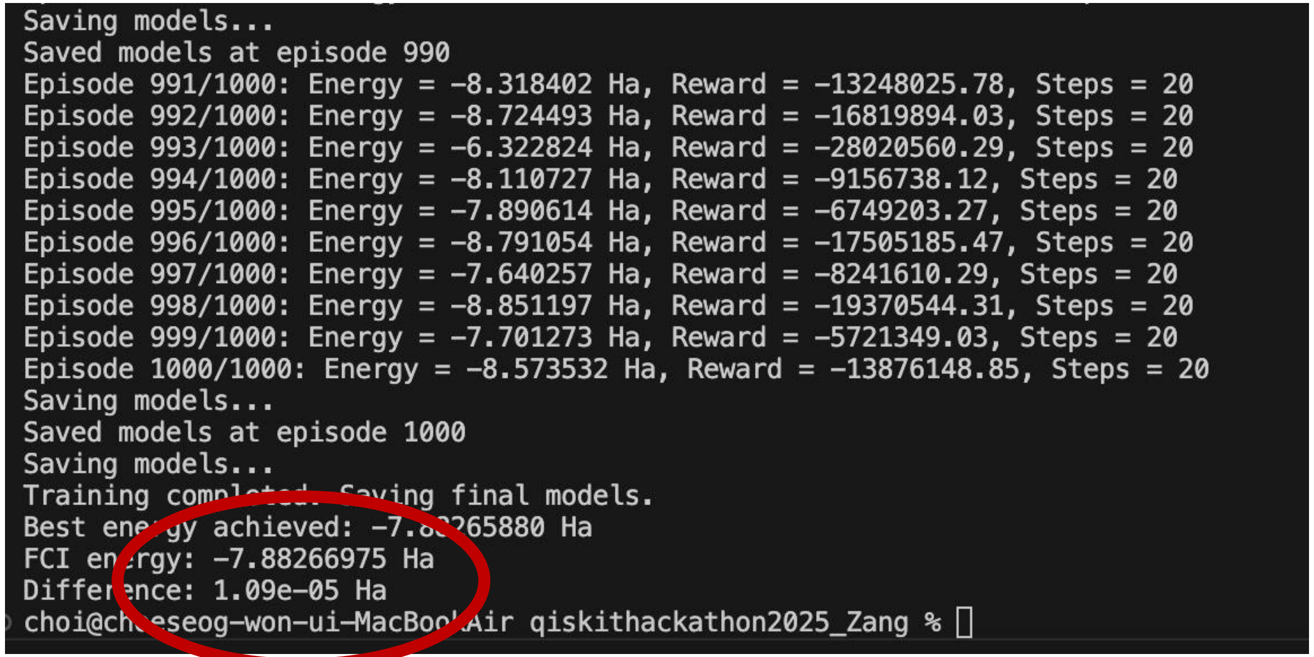
Noiseless Simulator
Provides an ideal, noise-free quantum environment to evaluate the baseline performance of quantum circuits.

```
Saving models...
Saved models at episode 990
Episode 991/1000: Energy = -7.698950 Ha, Reward = -202.79, Steps = 20
Episode 992/1000: Energy = -7.698950 Ha, Reward = -202.79, Steps = 20
Episode 993/1000: Energy = -7.698950 Ha, Reward = -385.60, Steps = 20
Episode 994/1000: Energy = -7.698950 Ha, Reward = -202.79, Steps = 20
Episode 995/1000: Energy = -7.698950 Ha, Reward = -202.79, Steps = 20
Episode 996/1000: Energy = -7.698950 Ha, Reward = -385.60, Steps = 20
Episode 997/1000: Energy = -7.698950 Ha, Reward = -202.79, Steps = 20
Episode 998/1000: Energy = -7.698950 Ha, Reward = -202.79, Steps = 20
Episode 999/1000: Energy = -7.698950 Ha, Reward = -202.79, Steps = 20
Episode 1000/1000: Energy = -7.698950 Ha, Reward = -202.79, Steps = 20
Saving models...
Saved models at episode 1000
Saving models...
Training completed. Saving final models.
Best energy achieved: -7.88134868 Ha
FCI energy: -7.88266975 Ha
Difference: 1.32e-03 Ha
choi@choeseog-won-ui-MacBookAir qiskithackathon2025_Zang %
```

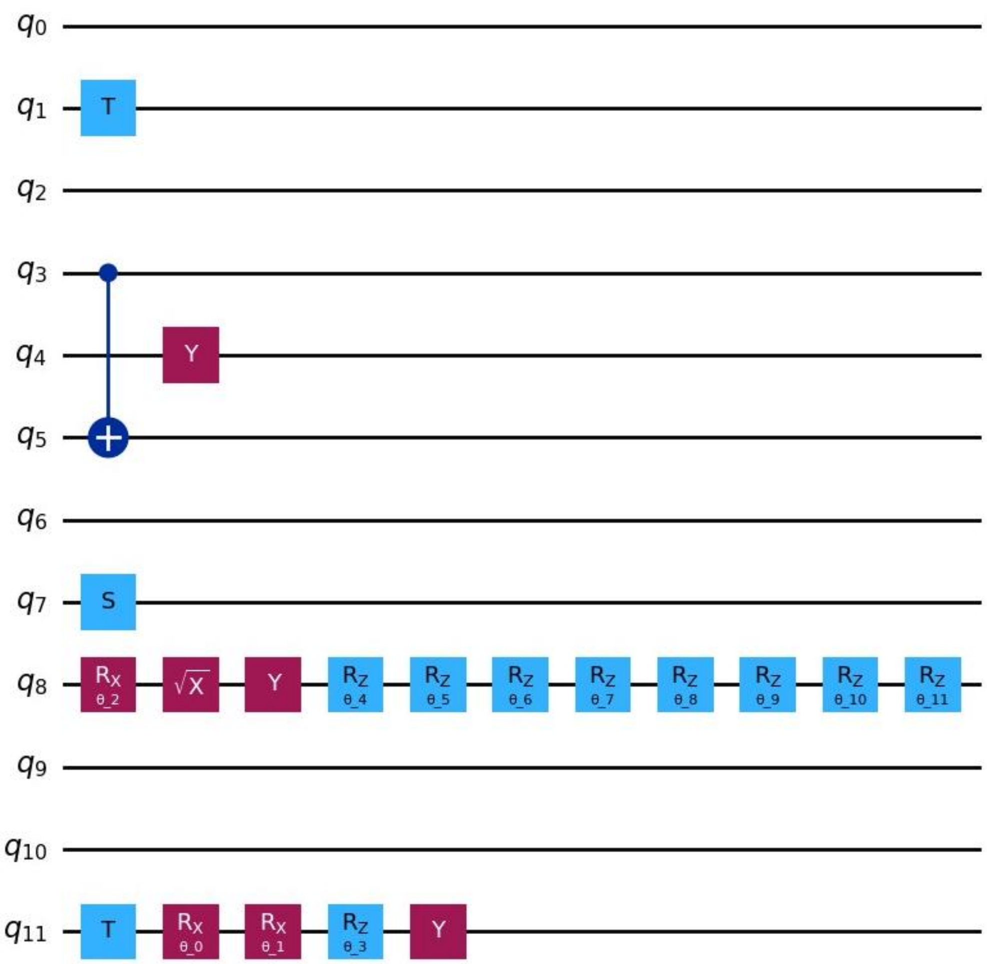


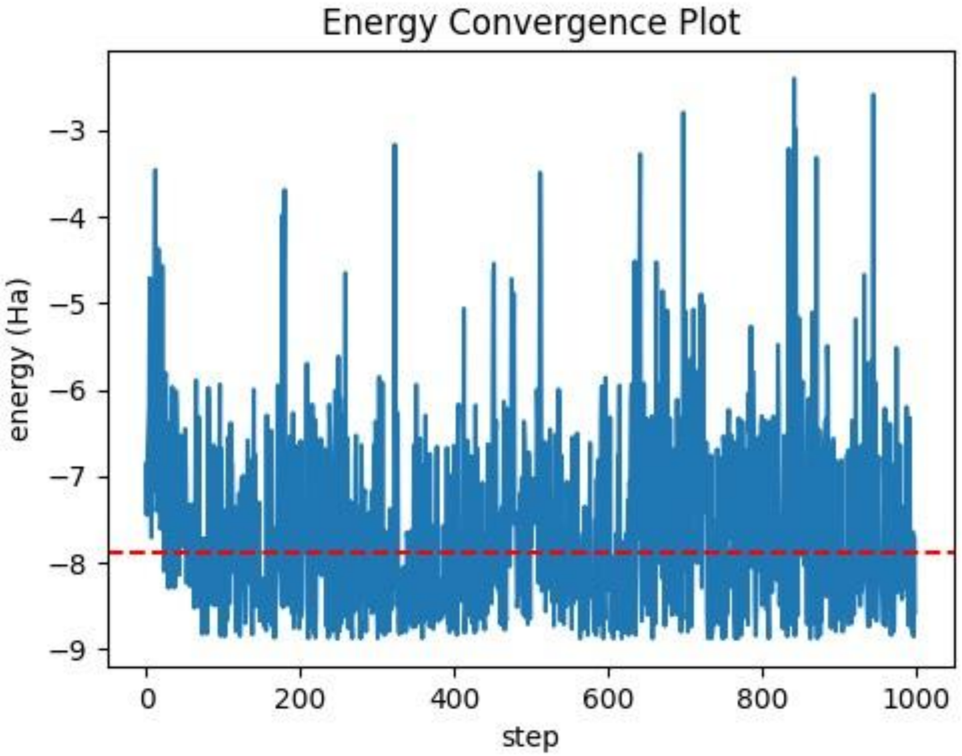


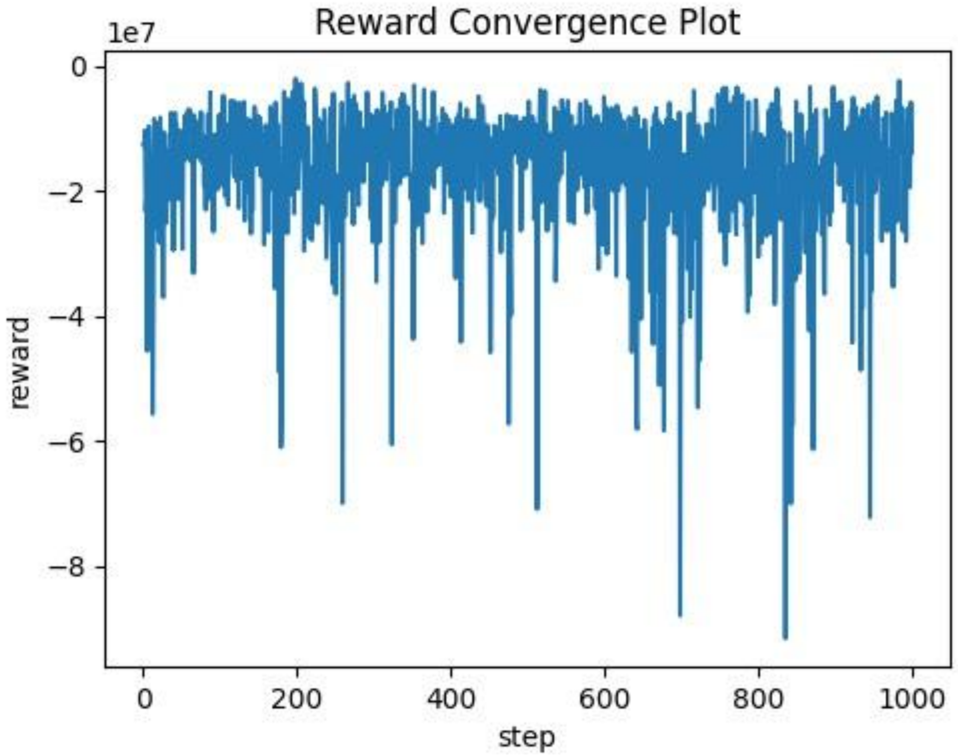
Energy Convergence Plot
energy (Ha)
−5
−6
−7
−8
−9
0
200
400
600
800
step
Reward Convergence Plot
reward
0
−10000
−20000
−30000
−40000
−50000
−60000
0
200
400
600
800
step



```
Saving models...
Saved models at episode 990
Episode 991/1000: Energy = -8.318402 Ha, Reward = -13248025.78, Steps = 20
Episode 992/1000: Energy = -8.724493 Ha, Reward = -16819894.03, Steps = 20
Episode 993/1000: Energy = -6.322824 Ha, Reward = -28020560.29, Steps = 20
Episode 994/1000: Energy = -8.110727 Ha, Reward = -9156738.12, Steps = 20
Episode 995/1000: Energy = -7.890614 Ha, Reward = -6749203.27, Steps = 20
Episode 996/1000: Energy = -8.791054 Ha, Reward = -17505185.47, Steps = 20
Episode 997/1000: Energy = -7.640257 Ha, Reward = -8241610.29, Steps = 20
Episode 998/1000: Energy = -8.851197 Ha, Reward = -19370544.31, Steps = 20
Episode 999/1000: Energy = -7.701273 Ha, Reward = -5721349.03, Steps = 20
Episode 1000/1000: Energy = -8.573532 Ha, Reward = -13876148.85, Steps = 20
Saving models...
Saved models at episode 1000
Saving models...
Training completed. Saving final models.
Best energy achieved: -7.88265880 Ha
FCI energy: -7.88266975 Ha
Difference: 1.09e-05 Ha
choi@cheeseog-won-ui-MacBookAir qiskithackathon2025_Zang %
```

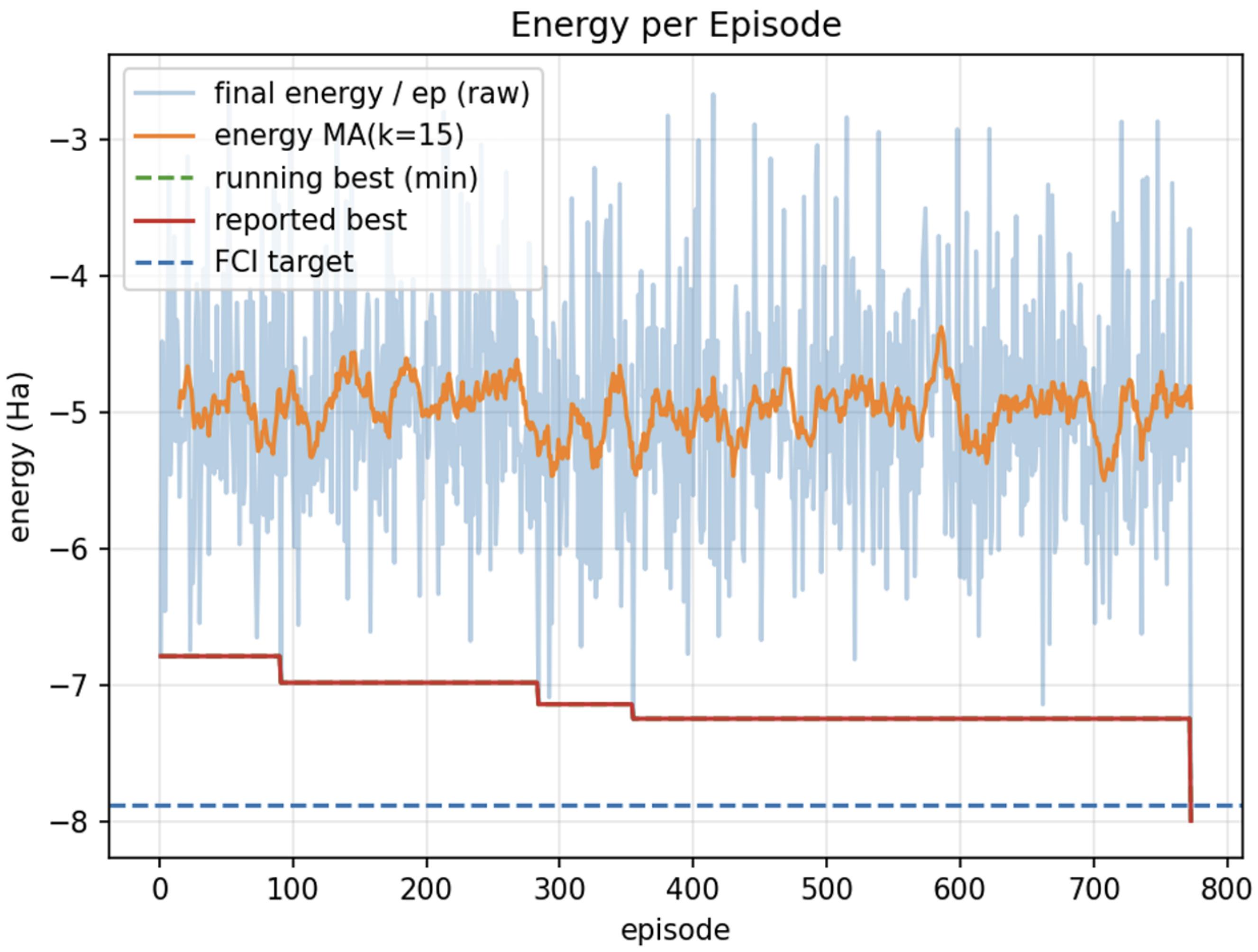



Energy Convergence Plot
energy (Ha)
−3
−4
−5
−6
−7
−8
−9
0
200
400
600
800
1000
step


Reward Convergence Plot
1e7
reward
0
−2
−4
−6
−8
0
200
400
600
800
1000
step

Experimental Results

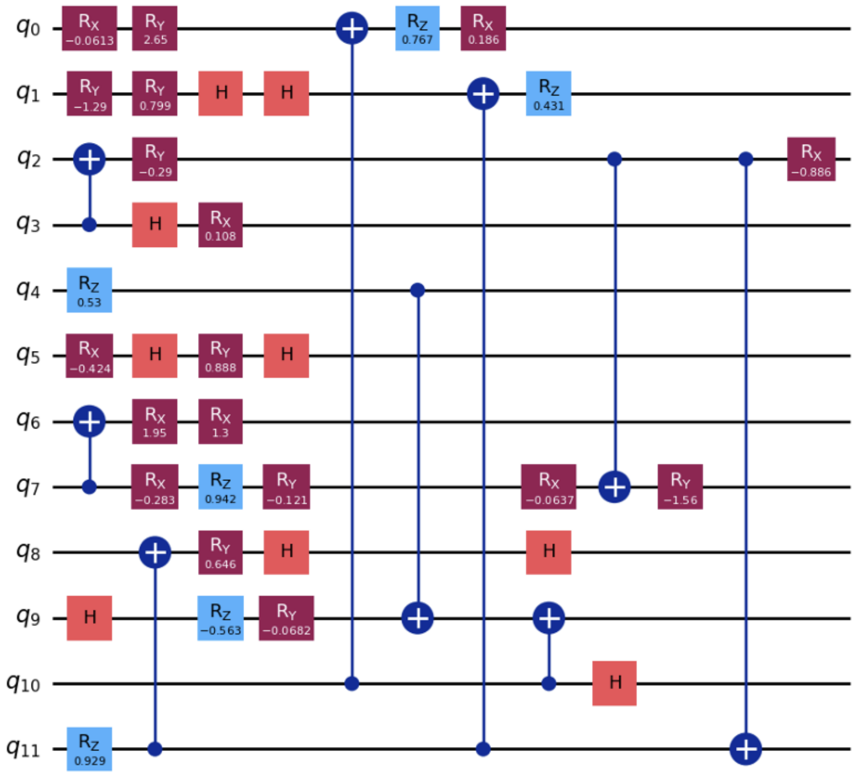
- LiH fci energy: -7.8827 Ha
- Converged at 780 episodes
- Final energy estimate
  -7.9953 Ha

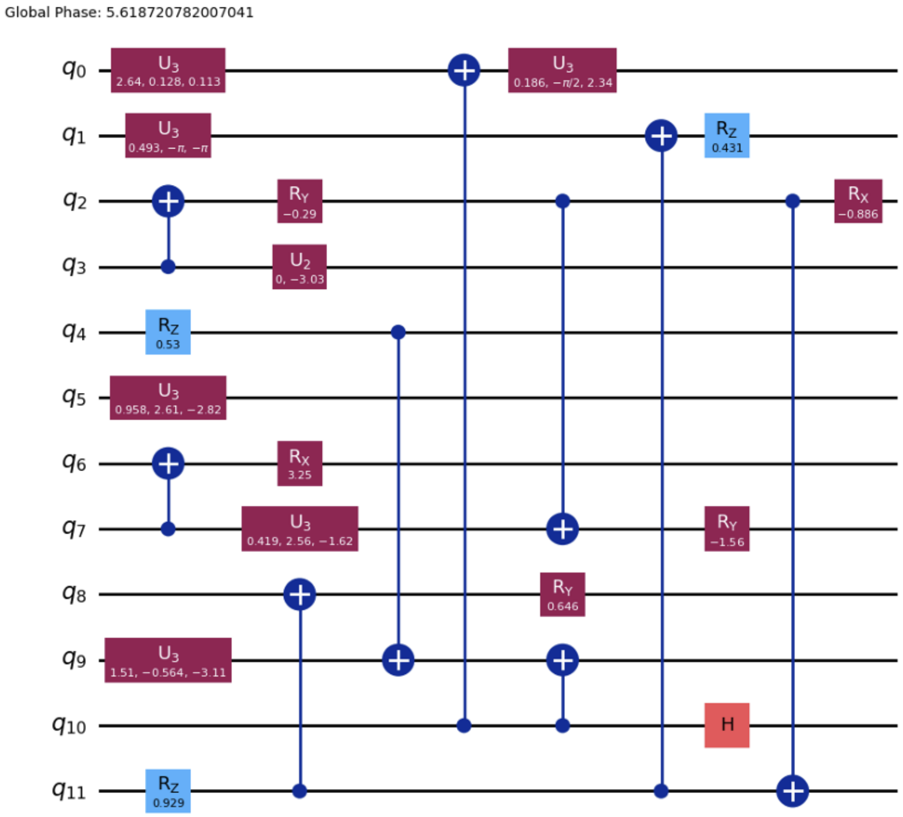


Experimental Results – Circuit

8 layers, 42 gates

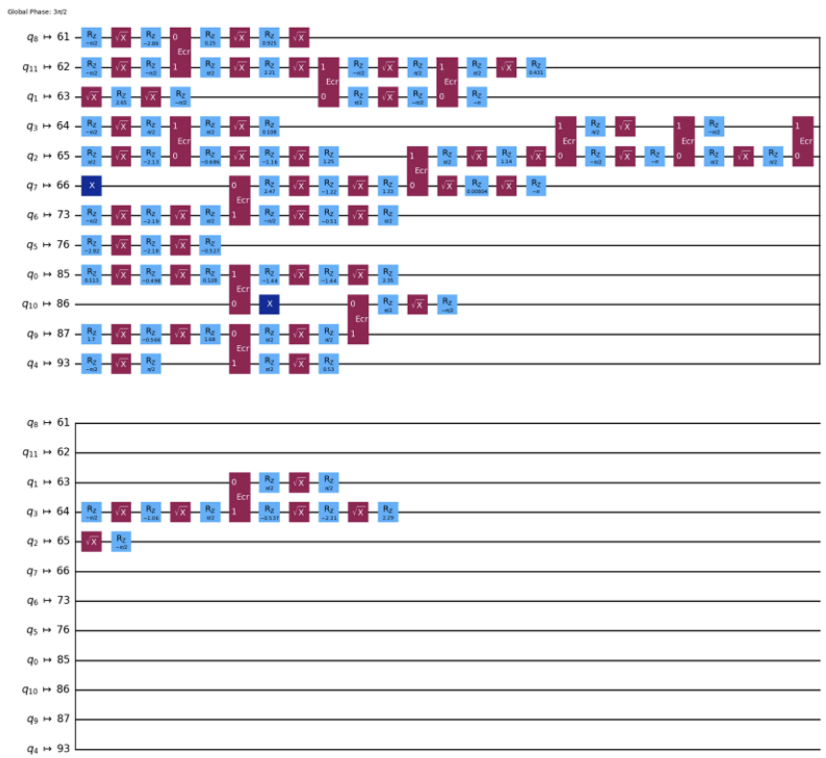
optimized to 5 layers, 25 gates



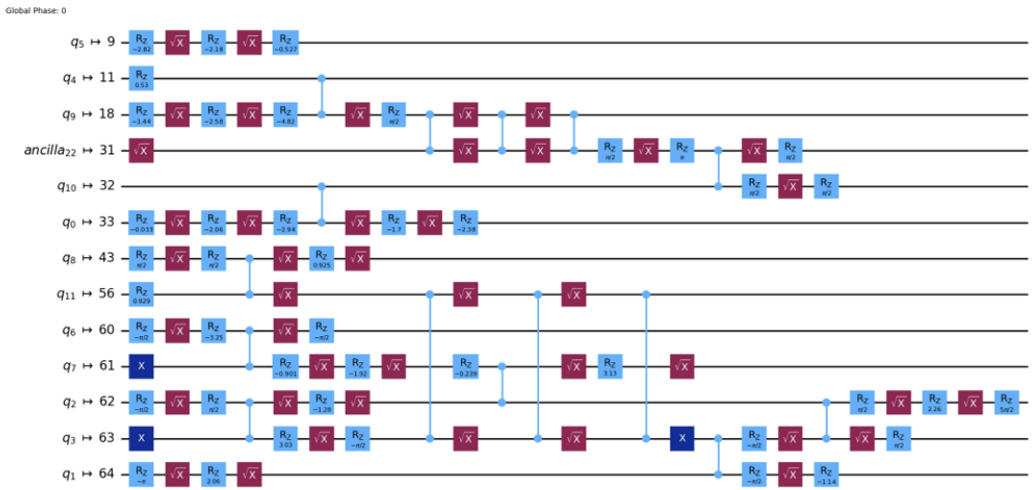


Experimental Results – Real Hardware

IBM Strasbourg (Eagle r3)
FCI Estimate: -7.6129 Ha



IBM Aachen (Heron r2)
FCI Estimate: -7.8894 Ha



Actual LiH FCI: -7.8827 Ha