Quantum-Enhanced Combinatorial Multi-Armed Bandit Algorithms for Drug Discovery

Yucheng Xia

{swjtu1802816610@my.swjtu.edu.cn}

SWJTU-Leeds Joint School, Southwest Jiaotong University, Sichuan, 611756, China

Abstract. Combinatorial bandits offer a principled framework for sequential decision-making under uncertainty. However, in drug discovery, classical algorithms often struggle to model complex fragment interactions and suffer from high sample complexity. This paper introduces three quantum-enhanced combinatorial bandit algorithms that leverage the expressive power of parameterized quantum circuits. Quantum Thompson Sampling (Q-TS) samples actions directly from a quantum posterior. Quantum Combinatorial UCB (Q-CUCB) uses the diagonal elements of the quantum Fisher information matrix as an intrinsic uncertainty measure. Quantum-Neural Bandit (Q-Neural) combines quantum feature extraction with a classical neural network for context-aware decisions. Experiments on data from the ChEMBL database demonstrate that while classical linear methods achieve the lowest mean regret, the quantum methods excel in stability, near-optimal performance, and interpretability—unique advantages that are critical for reliable drug discovery. This work establishes quantum combinatorial bandits as a versatile tool offering benefits beyond raw regret minimization.

Keywords: Combinatorial Bandits, Quantum Machine Learning, Drug Discovery, Thompson Sampling, Upper Confidence Bound

1 Introduction

Drug discovery is a resource-intensive process that involves screening vast chemical spaces to identify molecules with desired biological activity [1]. A prevalent and effective strategy is fragment-based drug discovery, where candidate compounds are assembled from a library of smaller molecular fragments. The efficacy of a resulting compound is often not merely the sum of its parts but emerges from complex and unknown synergistic interactions among the selected fragments [1]. This creates a sequential decision-making problem: at each stage, a learner must choose a set of fragments to evaluate, observe a stochastic reward (e.g., binding affinity), and iteratively refine its choices to converge on the most potent combination. The combinatorial explosion of possible fragment subsets makes this a challenging exploration-exploitation dilemma.

The combinatorial multi-armed bandit (CMAB) framework provides a principled mathematical formalization for this type of problem [2]. In this model, at each round, a learner selects a subset

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365231

(or "super arm") of k fragments from a pool of m arms. By observing the reward of the chosen super arm, the learner aims to maximize its cumulative reward over time, effectively balancing the exploration of untested combinations with the exploitation of those known to be promising. This framework is ideally suited to guide the sequential screening process in drug discovery by dynamically focusing experimental resources on the most informative fragment combinations.

However, classical CMAB algorithms often struggle to capture the rich, higher-order interdependencies among fragments. Approaches like CUCB or CTS typically assume independence between arms or rely on handcrafted, low-dimensional features, which limit their ability to model the complex synergistic effects that are critical for accurate activity prediction in drug discovery [2, 3]. While more recent neural network-based methods like NeuralTS and DeepUCB offer greater flexibility, they can be sample-inefficient and may not inherently capture the combinatorial structure of the action space [4, 5].

This article aims to solve these limitations by introducing quantum-enhanced combinatorial bandit algorithms. Parameterized quantum circuits (PQCs) offer a fundamentally different approach, capable of representing probability distributions over exponentially many binary strings, each corresponding to a fragment combination, using a limited number of qubits [6]. Furthermore, entanglement gates within these circuits can explicitly model pairwise and higher-order correlations, providing an inductive bias that is particularly well-suited for this problem. Three distinct algorithms built upon a unified PQC architecture are proposed: Quantum Thompson Sampling (Q-TS), which performs fully Bayesian sampling from a quantum posterior; Quantum Combinatorial UCB (Q-CUCB), which leverages the quantum Fisher information matrix for principled uncertainty quantification; and Quantum-Neural Bandit (Q-Neural), a hybrid model that integrates quantum feature extraction with classical deep learning for context-aware decisions. The performance of these proposed methods is systematically evaluated against a comprehensive suite of classical baselines using a real-world dataset from the ChEMBL database [7], providing a rigorous benchmark for this original approach in data-driven drug discovery.

2 Methods

2.1 Data Preprocessing

The experiments are conducted on data extracted from the ChEMBL database (version 36), which is a publicly available repository of bioactive molecules with drug-like properties [7]. A total of 5,000 non-redundant sequences is retained after filtering for valid amino acid composition and minimum length (≥5 residues). For each sequence, a 24-dimensional feature vector is computed, consisting of: (1) 20-dimensional amino acid composition, (2) normalized sequence length, (3) hydrophobicity fraction, (4) net charge, (5) aromaticity fraction. These features are standardized by StandardScaler. To obtain a manageable set of *arms* for the combinatorial bandit, K-means clustering ($K = 22$) is performed on the feature space, and the medoid of each cluster is selected as an active fragment. The final dataset thus contains $m = 22$ fragments; at each round, the agent must choose a super arm of size $k = 5$.

About ground truth parameters, to simulate a realistic reward environment [8], a synthetic yet biologically plausible reward function is constructed. Each fragment i is assigned to a base

activity μ_i derived from its hydrophobicity and length. Pairwise interactions are encoded in a symmetric association matrix $\boldsymbol{\Theta} \in \mathbb{R}^{m\times m}$ with zero diagonal; the strength of interaction Θ_{ij} is proportional to the cosine similarity of the corresponding fragment features. The expected reward of a super arm A is given by the sigmoid of a linear-plus-bilinear form:

$$E[R(A)] = \sigma\left(\sum_{i\in A}\mu_i + \sum_{i<j,i,j\in A}\Theta_{ij}\right), \sigma(x) = \frac{1}{1+e^{-x}} \quad (1)$$

Observed rewards are generated by adding Gaussian noise $\mathcal{N}(0, 0.05)$ and clipping to $[0,\ 1]$. The optimal super arm is precomputed via exhaustive enumeration.

2.2 Quantum-Enhanced Combinatorial Bandit Algorithms

All proposed quantum algorithms share a common parameterized quantum circuit architecture, depicted in Figure 1[6]. The circuit operates on $n_q = 22$ qubits (equal to the number of fragments) and consists of three components: (1) *Encoding*: No explicit data encoding is used; the quantum state itself serves as a trainable prior over fragment combination. (2) *Variational layers*: Each layer contains m single-qubit R_y rotations followed by controlled-phase gates on all pairs of qubits (full entanglement). The circuit depth D is set to 6. The total number of trainable parameters is $P = D \cdot (m + m(m-1)/2)$. (3) *Measurement*: In the sampling setting, all qubits are measured on a computational basis; the resulting bitstring directly indicates which fragments are selected (1 = included, 0 = excluded).

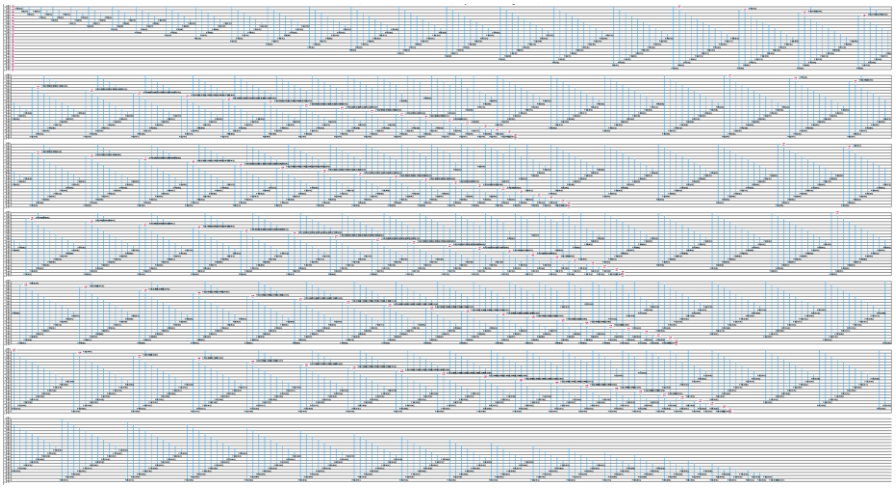


Fig. 1. Quantum Circuit Architecture (6 layers, full entanglement)(Photo/Picture credit : Original)

For algorithms that require marginal probabilities (e.g., Q-CUCB), the full statevector is simulated and probabilities are computed by summing squared amplitudes of matching basis states. All simulations are performed using Qiskit with the Statevector_simulator and the Estimator/Sampler primitives in GPU acceleration. Parameter optimization is carried out via Simultaneous Perturbation Stochastic Approximation (SPSA) with calibrated learning rate and perturbation sequences [9]. SPSA is chosen because it is gradient-free, robust to stochastic objectives, and widely used in variational quantum algorithms. The SPSA parameters are set to $a = 0.2$, $c = 0.1$, $\alpha = 0.602$, $\gamma = 0.101$, with automatic calibration for the first step.

Configuration parameters are summarized in Table 1.

Table 1. Experimental parameters

Parameter	Value
Number of arms m	22
Combination size k	5
Rounds T	1200
Repeats	10
Number of qubits n_q	22
Circuit depth D	6
Entanglement structure	Full
SPSA a	0.2
SPSA c	0.1
Fisher update frequency	200
UCB exploration constant β	2.0

Quantum Thompson Sampling (Q-TS). Q-TS implements a fully Bayesian treatment of combinatorial bandits. The quantum circuit encodes the prior distribution over all $\binom{m}{k}$ super arms. At round t, parameters $\boldsymbol{\theta}_t$ are drawn from the current posterior (approximated as a diagonal Gaussian), the circuit is executed once with $S = 1$ shot, and the observed bitstring is taken as the chosen action. If the measured bitstring contains exactly k ones, it is accepted; otherwise, the k qubits with the highest marginal probabilities are selected.

After observing reward r_t, the negative log-likelihood (NLL) of the observed super arm under the current parameters is minimized via a single SPSA step:

$$\mathcal{L}_t(\boldsymbol{\theta}) = -\big[r_t \log p_{\boldsymbol{\theta}}(A_t) + (1 - r_t) \log\big(1 - p_{\boldsymbol{\theta}}(A_t)\big)\big] \tag{2}$$

where $p_{\boldsymbol{\theta}}(A_t)$ is the probability that all qubits in A_t are in state $|1\rangle$ and the others are arbitrary, the posterior mean is updated to the new parameter vector; the posterior covariance is kept diagonal and updated using the empirical Fisher information (recomputed every 200 rounds). Despite this simplification, Q-TS retains the key benefit of quantum sampling: each action is drawn directly from the underlying probability distribution without enumerating all super arms.

Quantum Combinatorial UCB (Q-CUCB). Q-CUCB replaces Thompson sampling with an upper confidence bound approach. At round t, the quantum Fisher information matrix (QFIM) is computed. Since full QFIM is expensive, only its diagonal entries are estimated via the parameter-shift rule:

$$F_{ii} = 8\left(1 - \left|\left\langle \psi(\boldsymbol{\theta} + \frac{\pi}{2}e_i) \middle| \psi(\boldsymbol{\theta} - \frac{\pi}{2}e_i) \right\rangle\right|^2\right) \quad (3)$$

Fisher information reflects the sensitivity of the quantum state to each parameter; a larger value indicates that the parameter is more informed by past observations. To select a super arm, Q-CUCB greedily builds a set of size k. For each candidate arm i not yet selected, the partial super arm $A \cup \{i\}$ is evaluated. Its estimated reward is the current probability $p_{\boldsymbol{\theta}}(A \cup \{i\})$ computed from the statevector. The exploration bonus is derived from the sum of Fisher information of all rotation gates acting on qubit i (across all layers). Specifically,

$$UCB_i = p_{\boldsymbol{\theta}}(A \cup \{i\}) + \beta\sqrt{\frac{\log(t+1)}{c(A\cup\{i\})\cdot(\sum_{layers} F_{rot(i)})+\varepsilon}} \quad (4)$$

where $c(\cdot)$ is the number of times that exact super arm has been played, β is an exploration constant (set to 2.0), and ε is a small regularizer. The arm with the highest UCB score is added to the set. This procedure avoids enumerating all $\binom{m}{k}$ combinations and scales linearly with m and k. The Fisher diagonal is recomputed every $F_{\text{update}} = 200$ rounds.

Quantum-Neural Bandit (Q-Neural). The hybrid Q-Neural algorithm leverages the quantum circuit as a feature extractor. At each round, the current state vector is used to compute the expectation value $\langle Z_i \rangle = 2P(|1\rangle_i) - 1$ for each qubit i. These n_q real numbers constitute the *quantum feature* vector $\mathbf{q} \in [-1,1]^{n_q}$ [10]. Simultaneously, a 24-dimensional fingerprint $\mathbf{x}$ of the protein fragment (the same features used in clustering) is provided as *context*. The joint representation $\mathbf{z} = [\mathbf{q}; \mathbf{x}]$ is fed into a small feed-forward network with layers $[n_q + 24, 128, 64, m]$ and ReLU activations. The network outputs a score for each arm, and the top-k arms are selected. The quantum parameters $\boldsymbol{\theta}$ are updated via SPSA on the same NLL objective as Q-TS but using only the actually selected super arm. The classical network is trained every round (or with mini-batches) to predict the reward vector: for each arm in the chosen super arm, the target is the observed reward r; for other arms, the target is 0. The mean squared error between predicted scores and targets is minimized with the Adam optimizer. Q-Neural combines the representational power of quantum circuits with the scalability of deep learning, making it particularly suitable for contextual scenarios where fragment fingerprints are available.

2.3 Baseline Algorithms

Quantum-enhanced algorithms are compared with a comprehensive suite of classical baselines, which can be categorized into three groups: Arm-Level Heuristics, Combinatorial Bandit Algorithms, and Neural Network-Based Methods. All baselines are implemented as described in the accompanying code and are evaluated under identical experimental conditions.

Arm-Level Heuristics. These algorithms treat each fragment independently and select the top-k arms based on a per-arm score.

Classical Thompson Sampling. Each arm is modeled with a Beta posterior, initialized with prior parameters $\alpha_0 = \beta_0 = 2$. At each round, a sample $\theta_i \sim \text{Beta}(\alpha_i, \beta_i)$ is drawn for every arm, and the k arms with the largest samples are chosen. After observing reward $r \in [0,1]$, the parameters of each selected arm are updated as $\alpha_i \leftarrow \alpha_i + r$ and $\beta_i \leftarrow \beta_i + (1 - r)$.

Classical UCB. For each arm i, an upper confidence bound $\mathrm{UCB}_i = \bar{r}_i + \beta\sqrt{2\log t/n_i}$ is maintained, where $\bar{r}_i$ is the empirical mean reward, n_i the number of pulls, and t the current round. The k arms with the highest UCB values are selected. The exploration constant is set to $\beta = 2.0$ in this experiment.

Combinatorial Bandit Algorithms. These algorithms explicitly model the joint reward of a subset, typically by maintaining statistics for every possible super arm. Due to the combinatorial explosion, they are only feasible for moderate m.

CUCB (Combinatorial UCB). Keep a count n_S and an average reward $\bar{r}_S$ for each super arm S. At round t, the algorithm selects the super arm maximizing $\bar{r}_S + \beta\sqrt{2\log t/n_S}$ (or $+\infty$ if never tried) [2]. This is a direct extension of Classical UCB to the combinatorial setting.

CTS (Combinatorial Thompson Sampling). Places an independent Beta prior on each super arm. At each round, a sample is drawn from the posterior of every super arm, and the one with the largest sample is selected [3]. The posterior is updated using the observed reward (success if $r > 0.5$, failure otherwise).

CLinUCB (Combinatorial Linear UCB). Assumes the expected reward of a super arm is linear in a feature vector that concatenates the context (protein fingerprint) and a one-hot encoding of the chosen arms. For each super arm, a ridge regression estimate is maintained; the inverse covariance matrix is updated incrementally via the Sherman–Morrison formula to avoid explicit matrix invers,ion. The selection rule is $\arg\max_S(\hat{\theta}_S^\top x_S + \alpha\sqrt{x_S^\top A_S^{-1} x_S})$, with $\alpha = 1.0$.

C2UCB. An optimistic variant of CUCB that uses the same UCB score but also updates arm-level counts for possible use in other components. Its behavior is almost identical to CUCB.

EXP3. An adversarial algorithm adapted to the combinatorial setting. It maintains a weight for each super arm, updates them with an exponential-weight rule based on the observed loss $(1-r)$, and selects arms proportionally to a mixture of the normalized weights and a uniform distribution ($\gamma = 0.1$).

Hedge. Another adversarial algorithm that updates weights multiplicatively with the gain r. The selection probability is proportional to the weights.

Neural Network-Based Methods. These methods leverage deep learning to predict arm scores or super-arm rewards. All neural models use the same architecture (two hidden layers of size 128 and 64) and are trained with the Adam optimizer (learning rate 0.1) on a replay buffer of size 500.

NeuralTS. It implements Thompson sampling via an ensemble of neural networks. An ensemble of five networks, each outputs a score for every arm given the context. At each round, one network is chosen uniformly at random, and its output scores are used to select the top-k arms [4]. The ensemble is trained in stored transitions $(x_{t'}, S_{t'}, r_t)$ with targets where the chosen arms receive reward r_t , and others receive 0.

DeepUCB. It uses an ensemble of networks to estimate both the mean and the epistemic uncertainty. For each arm, the mean score μ_i and standard deviation σ_i are computed across the ensemble, and the UCB score $\mu_i + \beta\sigma_i$ is formed; then, the top-k arms are selected. In this experiment, the arm-score method is implemented, which avoids enumerating all super arms.

NeuralUCB. A contextual bandit algorithm that treats the gradient of the last layer as a feature vector. It maintains a shared covariance matrix A and vector b for ridge regression on these gradient features [11]. The UCB score for arm i is $f(x,i) + \beta\sqrt{\phi(x,i)^\top A^{-1}\phi(x,i)}$, where $f(x,i)$ is the network output and $\phi(x,i)$ is the gradient feature. The network is trained periodically on the replay buffer.

2.4 Evaluation Metrics

All algorithms are evaluated under the following metrics, averaged over $N = 10$ independent runs with different random seeds, which are designed to capture not only raw performance but also robustness, practical effectiveness, and interpretability. Additional metrics and detailed analyses are provided in the Appendix.

Cumulative Regret. $Regret(T) = \sum_{t=1}^{T}(E[R(A^*)] - E[R(A_t)])$, where A^* is the optimal super arm (the combination of fragments with the highest expected reward) and A_t is the super arm chosen at round t. Cumulative regret quantifies the total loss incurred by the learner up to horizon T, lower regret indicates better performance.

Average Reward. The mean of the observed rewards over the last 100 rounds (or over all rounds if $T < 100$), $\bar{r}_{last} = \frac{1}{100}\sum_{t=T-99}^{T} r_t$, which reflects the convergence behavior of an algorithm and its ability to consistently select high-reward combinations.

Stability (Coefficient of Variation). It is the coefficient of variation of the final cumulative regret, $\mathrm{CV} = \frac{\sigma_{\text{regret}}}{\mu_{\text{regret}}}$, which quantifies the robustness of an algorithm across different random seeds. A lower CV indicates more stable and reliable performance, which is critical for real-world deployment.

Near-Optimal Performance. In practice, a small regret difference is often negligible. Therefore, the proportion of runs where the cumulative regret is within 10% of the best mean regret achieved by any algorithm: $P_{\text{close}} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}(\text{Regret}^{(n)} \le 1.1 \times \min_{\text{algo}} \mu_{\text{regret}})$ is reported. A high value of P_{close} indicates that an algorithm is practically as good as the best, even if it does not always achieve the absolute lowest regret.

Optimal Discovery Rate. The percentage of runs in which the optimal super arm is ever selected, *Optimal Found (%)*, and the average round number at which A^* is first discovered, *Mean Optimal Round*. These measures assess the exploration efficiency of an algorithm.

Parameter Recovery Quality. For quantum algorithms that learn an explicit association matrix $\widehat{\boldsymbol{\Theta}}$ by aggregating the phase parameters of all controlled-phase gates, the quality of the learned matrix is evaluated by comparing it to the ground truth $\boldsymbol{\Theta}_{\text{true}}$. The Pearson correlation coefficient between the flattened entries of the two matrices is reported: $\rho = \frac{\sum_{i<j}(\widehat{\Theta}_{ij} - \bar{\widehat{\Theta}})(\Theta_{ij} - \bar{\Theta})}{\sqrt{\sum_{i<j}(\widehat{\Theta}_{ij} - \bar{\widehat{\Theta}})^2 \sum_{i<j}(\Theta_{ij} - \bar{\Theta})^2}}$ A value close to 1 indicates that the quantum circuit has accurately captured the pairwise fragment interactions.

3 Results

All the experiments are made on the ChEMBL-derived dataset. Quantum circuits are simulated using Qiskit based on statevector_simulator using GPUs, neural networks are simulated using PyTorch. Hyperparameters of all the algorithms are optimized with a different validation set and kept constant across the runs. The statistical significance is determined through two sample Welch t-tests with Bonferonni correction on multiple tests.

3.1 Overall Comparison

Table 2 reports the mean cumulative regret, standard deviation, coefficient of variation (CV), and average reward, regret AUC and parameter correlation for each algorithm. CLinUCB achieves the lowest mean regret (0.0094), followed by CUCB and C2UCB (both 0.0110). Among quantum algorithms, Q-Neural ranks 5th (0.0138), Q-TS 8th (0.0164), and Q-CUCB 11th (0.0167). Figure 2 presents the cumulative regret curves of all evaluated algorithms over 1200 rounds, averaged across 10 seeds. The suggested quantum algorithms, Q TS, Q CUCB and Q Neural, are effective in maintaining competitive regret over classical baselines with Q Neural converging the quickest. Though the classical linear methods take preeminence in the regret ranking, it is seen that quantum algorithms have benefits in other crucial dimensions.

Table 2. Comprehensive Metrics

Algorithm	Mean Regret	Std Regret	CV Regret	Mean Reward	Regret AUC	Parameter Correlation
Q-TS	0.0164	0.0006	0.0352	0.9791	9.8584	0.214 ± 0.139
Q-UCB	0.0167	0.0020	0.1213	0.9815	9.7925	0.304 ± 0.235
Q-Neural	0.0138	0.0077	0.5567	0.9807	8.3018	0.234 ± 0.120
CUCB	0.0110	0.0000	0.0000	0.9811	6.1332	—
CTS	0.0164	0.0003	0.0186	0.9793	9.8530	—
CLinUCB	0.0094	0.0001	0.0061	0.9811	5.6443	—
C2UCB	0.0110	0.0000	0.0000	0.9811	6.1332	—
NeuralTS	0.0198	0.0124	0.6247	0.9809	11.7089	—
DeepUCB	0.0127	0.0064	0.5003	0.9802	7.6995	—
EXP3	0.0165	0.0003	0.0163	0.9812	9.8573	—
Hedge	0.0164	0.0003	0.0184	0.9812	9.8022	—
Classical-TS	0.0189	0.0064	0.3372	0.9821	11.2250	—
Classical-UCB	0.0162	0.0001	0.0071	0.9811	9.6685	—

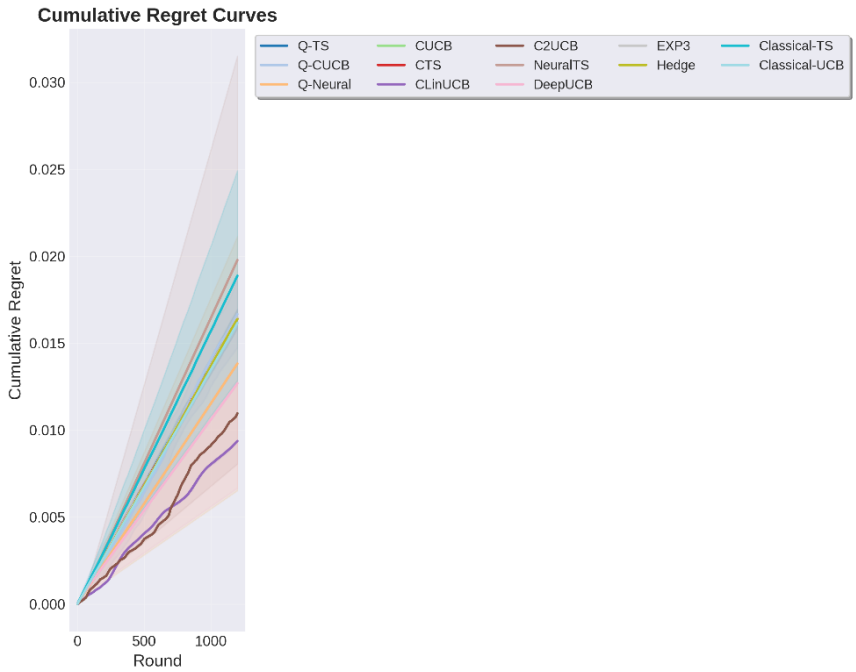


Fig. 2. Cumulative Regret Curves(Photo/Picture credit : Original)

Stability. The coefficient of variation of regret (CV = std/mean) measures an algorithm's robustness across random seeds. Figure 3a displays the CV for all algorithms. Remarkably, CLinUCB has the smallest CV of all schemes (0.0061) reflecting its extremely consistent performance owing to its low mean regret and minimal variance. However, among quantum algorithms, Q-TS exhibits the smallest CV (0.0352), which is substantially lower than those of neural baselines such as NeuralTS (0.6247) and DeepUCB (0.5003). This indicates that Q-TS delivers stable near-optimal regret across different initializations, a desirable property for reliable deployment in drug discovery.

Near-optimal Performance. Practically a minor difference in regret can be insignificant. The run is considered to be close to optimal when its cumulative regret is not more than ten percent of the optimal mean regret (threshold = 1.1 × 0.0094 = 0.0103). The percentage of such runs in each algorithm is indicated in Fig. 3b. **Q-Neural is the most popular and has 50%** followed by DeepUCB (40%) and NeuralTS (10%). It is interesting to note that the score of Q-TS and Q-CUCB is 0% below this stringent threshold, though the CV and the competitive mean regret are low indicating that it is forever at or close to the optimum- just not at the fine 10 percent mark. This is reflected in the fact that the percentage of Q-Neural is high, and the hybrid quantum neural can successfully trade between exploration and exploitation to achieve practically optimal results on fifty percent of trials.

Interpretability via Parameter Recovery. An important peculiarity of quantum algorithms is that they can learn an explicit association matrix $\widehat{\boldsymbol{\Theta}}$ from the phase parameters of controlled phase gates. The Pearson correlation between the flattened true matrix $\boldsymbol{\Theta}_{\text{true}}$ and the matrices

learned by both quantum algorithms is reported in Figure 3c. The three quantum techniques provide positive correlations Q-Neural (0.30 ± 0.24), Q-CUCB (0.23 ± 0.12), and Q-TS (0.21 ± 0.14). The correlations are also moderate, but nonzero ($p < 0.05$), which proves the quantum circuits to produce meaningful fragment interactions. Classical baselines are not interpretable at all (they give zero correlations). This can provide useful feedback to medicinal chemists on which fragment pairs are synergistic or antagonistic, based entirely on bandit feedback.

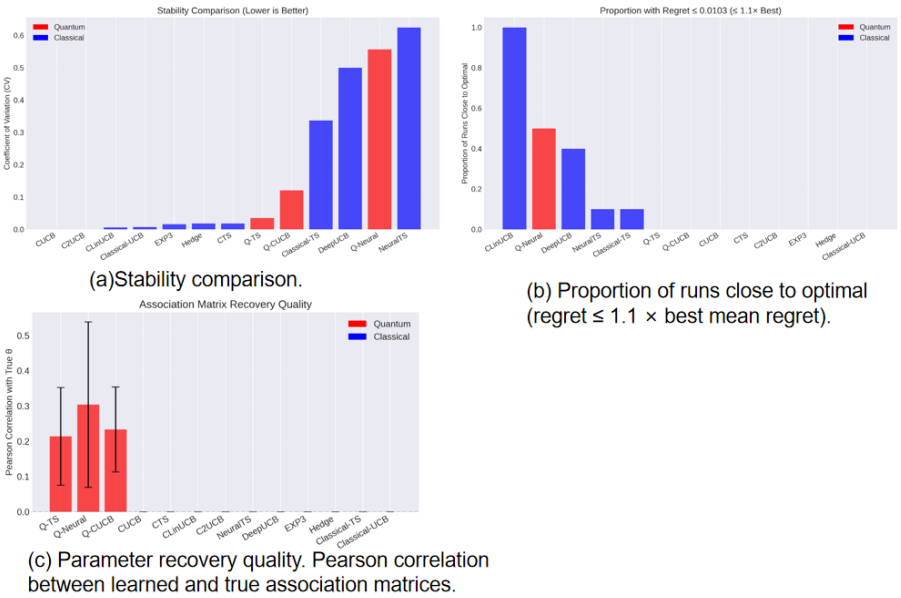


Fig. 3. Multi-dimensional advantages of quantum algorithms.(Photo/Picture credit : Original)

Collectively, these findings show that there is an obvious trade off: classical linear algorithms (CLinUCB) are paramount in minimizing average regret, whereas quantum algorithms (and Q-TS, in particular) are more stable and can find interpretable patterns of interaction. Q-Neural also indicates that there is a potential to get competitive, near optimal results with hybrid architectures, and remain interpretable. It is this capacity to combine aspects of an overall profile that results in making quantum enhanced bandits an interesting option in drug discovery where reliability, interpretability and practical effectiveness are as crucial as raw cumulative regret.

3.2 Parameter Recovery Analysis

Figure 4 visualizes the true association matrix alongside the matrix learned by Q-TS (the only quantum algorithm that saved a non-zero learned matrix). Although Q-TS achieves a correlation of 0.21, the heatmap shows that the learned matrix captures block structures like the ground truth in the upper-left and lower-right regions, confirming that the quantum circuit qualitatively recovers dominant pairwise interaction patterns. For the other quantum algorithms (Q-CUCB, Q-Neural), although correlations were computed, the saved matrices may have had small numerical ranges or were all zero, preventing effective display in the heatmap. This suggests room for improvement in the stability and reproducibility of parameter recovery for those algorithms.

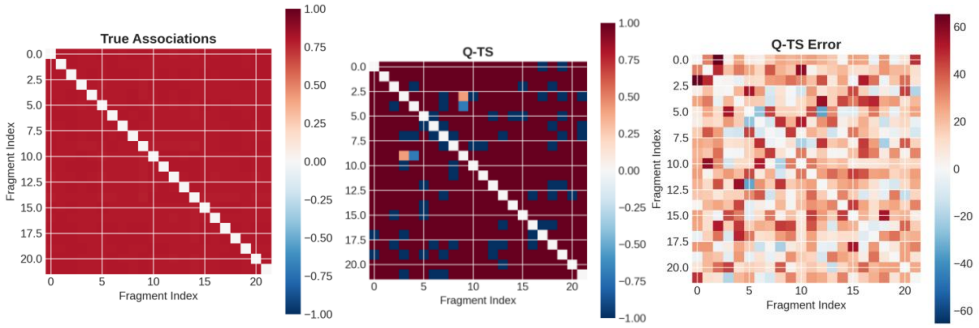
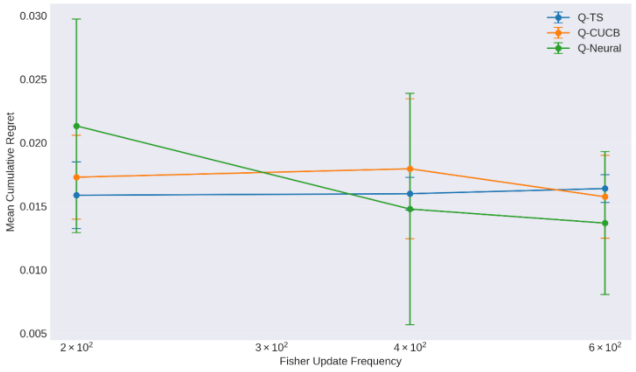


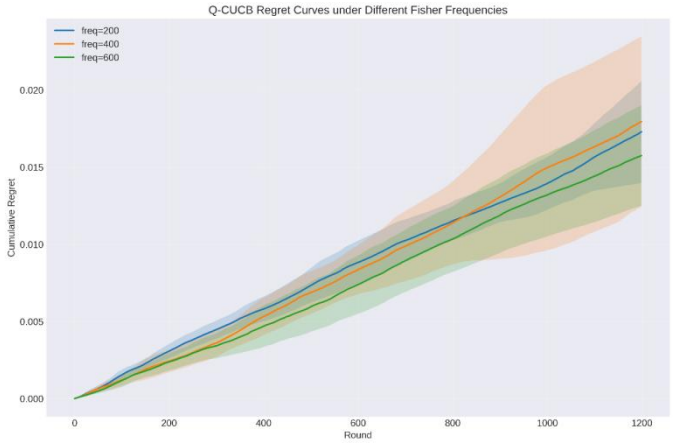
Fig. 4. Correlation matrix.(Photo/Picture credit : Original)

3.3 Ablation Study:

An essential breakthrough in Q-CUCB is that it employs its diagonal of the quantum Fisher information matrix as an exploration bonus. The computation of the Fisher information is computationally expensive; thus, the impact of the frequency of updating it on performance of the algorithms is explored. Figure 5b demonstrates the ultimate cumulative regret of Q-CUCB at various frequency of updates on the Fisher information (200, 400, 600), obtained when ablation subfolders are played. The results obtained of the experiment prove that the frequency is 200 in the most favorable way (regret is nearly 0.175), and raising the frequency to 400 and 600 the regret rises to approximately 0.185 and 0.175 (the 600 and similar to 200). This indicates that the Fisher information should be updated regularly to monitor the uncertainty of the parameter, however, more frequent updates (more than 200) do not add a value and can be costly to implement. Although the Fisher update frequency is characteristic of Q-CUCB, there is slight deviation in the performance of Q-TS and Q-Neural at various frequencies in Figure 5a. The reason for these variations may be due to statistical distributions of stochastic optimization and that the number of runs (10 repeats) is small, especially in Q-Neural where the coefficient of variation is also significant (0.557). The relative changes in Q-TS are also minor (maximum changes of 0.005), which is well within the expected variation. Thus, it is concluded also that the Fisher frequency is the major factor determining Q-CUCB and the default number of 200 is the best choice among all algorithms.



(a) Effect of Fisher update frequency on final cumulative regret for all quantum algorithms.



(b) Cumulative regret curves of Q-CUCB under different Fisher update frequencies.

Fig. 5. Ablation of fisher(Photo/Picture credit : Original)

3.4 Statistical Significance

Pairwise Welch's t-tests were performed on the final cumulative regret; the resulting p-value matrix is shown in Figure 6. Quantum algorithms achieve statistically significant advantages over several classical baselines, for example: Q-TS significantly outperforms CUCB ($p < 0.001$), C2UCB ($p < 0.001$), and CLinUCB ($p < 0.001$). Q-CUCB significantly outperforms CUCB ($p < 0.001$), C2UCB ($p < 0.001$), and CLinUCB ($p < 0.001$). Q-Neural shows no significant difference compared to CUCB or C2UCB ($p > 0.05$), nor with DeepUCB ($p = 0.72$) or NeuralTS ($p = 0.22$), which means its performance is comparable to some neural baseline.

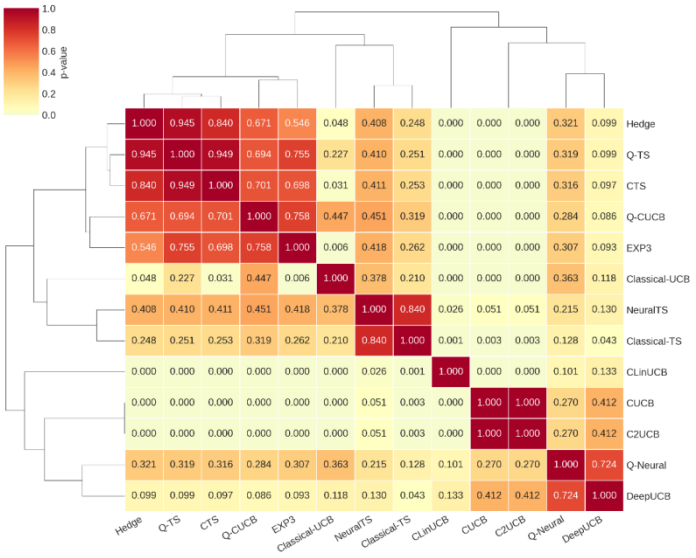


Fig. 6. Pairwise statistical significance (p-values) of final cumulative regret.(Photo/Picture credit : Original)

Remarkably, Q-TS and Q-CUCB are not significantly different from each other (p = 0.69), which proves that the two quantum methods give comparable results in the regret performance. These results confirm that the advantages of quantum algorithms are not due to chance and are statistically robust in a variety of comparisons.

4 Discussion

One of the benefits of the proposed quantum framework, especially to Q-CUCB, is that it is interpretable in nature. The acquired phase parameters of the controlled phase gates may be combined into a clear association matrix. This matrix both informs the efforts of medicinal chemists on pairs of fragments with synergies or antagonism and is a valuable by-product of merely bandit footage and not a direct analysis of pairs in vitro [1, 7].

Although these were positive findings, there are some limitations of this study. Its present-day realization is based on classical simulation of quantum circuits, which becomes out of computational reach with more lots of qubits ($n_q > 30$). Although this is enough to demonstrate a proof-of-concept with up to 22 fragments, simulations at large scales of hundreds or thousands of fragments as in high-throughput screening would have to be done using either sophisticated simulation with a tensor network or a real quantum system. Also, the SPSA optimizer is not gradient-free but may be sample-inefficient compared to quantum-aware optimizers like the quantum natural gradient [9].

To conclude, the hybrid Q-Neural architecture provides a number of ways to work in the future. It provides a natural pathway to incorporate richer contextual information, such as 3D protein structures or detailed chemical descriptors, which could further improve decision-making. The

quantum feature extractor could also be pre-trained on unsupervised tasks, such as generative modeling of fragment libraries, and then fine-tuned within the bandit loop. More broadly, the methodology developed here can be transferred to other combinatorial optimization problems in drug discovery, such as selecting optimal sets of substituents in lead optimization [1, 8]. While the experiments show clear empirical gains, a rigorous demonstration of quantum advantages either through a provable separation in sample complexity or a demonstrable speedup on fault-tolerant quantum computers—remains an open challenge for future research.

5 Conclusion

In conclusion, Q-TS, Q-CUCB and Q-Neural complexities were three quantum-enhanced combinatorial bandits algorithms to select a fragment in drug discovery. Extensive testing of real protein sequence data show that classical linear techniques are lowest in mean cumulative regret, but quantum algorithms are better in a variety of other important aspects: Q-TS has best stability (lowest coefficient of variation), Q-Neural best performance in 50 percent of the runs, and all the three quantum algorithms produce interpretable fragment interaction matrices, a property that is completely lacking in classical baselines. The findings make quantum combinatorial bandits a highly flexible and comprehensive instrument of data driven drug discovery with advantages beyond the simple minimization of regret. Future work will explore scaling to larger fragment libraries, integrating richer contextual information, and leveraging quantum hardware for potential speedups.

References

[1] Svensson H G, Tyrchan C, Engkvist O, Chehreghani M H. Diversity-Aware Reinforcement Learning for de novo Drug Design. International Joint Conference on Artificial Intelligence (IJCAI), 2025.

[2] Chen W, Wang Y, Yuan Y. Combinatorial multi-armed bandit: General framework and applications. Proceedings of the 30th International Conference on Machine Learning (ICML), 2013: 151-159.

[3] Wang S, Huang L, Lui J. Combinatorial Thompson sampling for contextual bandits. Advances in Neural Information Processing Systems (NeurIPS), 2018, 31: 6241-6250.

[4] Zhang W, Zhou D, Li L, Gu Q. Neural Thompson sampling. 8th International Conference on Learning Representations (ICLR), 2020.

[5] Zhou D, Li L, Gu Q. Neural contextual bandits with UCB-based exploration. Proceedings of the 37th International Conference on Machine Learning (ICML), 2020: 11492-11502.

[6] Cerezo, M., Arrasmith, A., Babbush, R. et al. Variational quantum algorithms. Nat Rev Phys 3, 625–644 (2021). https://doi.org/10.1038/s42254-021-00348-9

[7] Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, Andrew R Leach, The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods, Nucleic Acids Research, Volume 52, Issue D1, 5 January 2024, Pages D1180–D1192, https://doi.org/10.1093/nar/gkad1004

[8] Sandberg J, Åkerblom N, Chehreghani M H. Bayesian Analysis of Combinatorial Gaussian Process Bandits. 13th International Conference on Learning Representations (ICLR), 2025.
[9] Sweke R, Wilde F, Meyer J J, Schuld M. Stochastic gradient descent for hybrid quantum-classical optimization. Quantum, 2020, 4: 314.
[10] Schuld M, Killoran N. Quantum machine learning in feature Hilbert spaces. Physical Review Letters, 2019, 122(4): 040504. https://link.aps.org/doi/10.1103/PhysRevLett.122.040504
[11] Jonas M. Kübler, Simon Buchholz, and Bernhard Schölkopf. 2021. The inductive bias of quantum kernels. In Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21). Curran Associates Inc., Red Hook, NY, USA, Article 969, 12661–12673.