ISSN: 2093-5374 / E-ISSN: 2093-5382

Quantum-Inspired Reinforcement Learning for Adaptive Semiconductor Wafer Probing in Multi-Die Environments

Srinivasa Rao Gondi[1*]

[1*]Senior Principal Test Engineer, NXP Semiconductors San Jose, California, USA.
gondi.srini@gmail.com, https://orcid.org/0009-0009-8853-6056

Received: February 14, 2026; Revised: March 21, 2026; Accepted: May 08, 2026; Published: June 30, 2026

Abstract

It has been suggested in the present paper that a Quantum-Inspired Reinforcement Learning (QIRL) framework can be proposed to adaptive semiconductor wafer probing in intricate multi-die settings, where traditional deterministic policies and classical reinforcement learning frameworks have difficulties in combinatorial explosion, sparse reward gradients, and uneven distributions of defects. The proposed approach introduces an amplitude-based superposition mechanism for probabilistic policy encoding, enabling the simultaneous exploration of multiple probing trajectories and improving decision-making efficiency under uncertainty. The environment of the high-fidelity digital twin simulation was created based on the realistic conditions of the wafer, which included the heterogeneous die layout, stochastic defect patterns (random, clustered and edge-based), and probe degradation dynamics. QIRL framework was strictly tested with comparison to the baseline procedures that were deterministic raster scanning, tabular Q-learning, Deep Q-Network (DQN), and Proximal Policy Optimization (PPO). The effectiveness of the proposed model is shown to be reduced to 28 % less probing moves than raster scanning and 17 % less than classical Q-learning through experimentation. Also, the switching costs are minimized up to 22% compared to PPO, which means an increase in the efficiency of path optimization. QIRL has a defect detection accuracy of 94/97 on average with a false positive rate of less than 3 meaning that in clustered defect situations, QIRL is able to find defects with a high accuracy of up to 97. Moreover, the framework also decreases the wear of probes by about 20 %, which leads to the increased lifespan of the equipment and reliability of the testing. All in all, the presented QIRL model is an excellent balance between the probing efficiency, accuracy of defect localization, and hardware durability. The findings indicate that it is scalable and can be used in next-generation semiconductor testing, and its potential is high when it comes to implementation in hybrid quantum-classical computing systems.

Keywords: Quantum-Inspired Reinforcement Learning (QIRL), Semiconductor Wafer Probing, Multi-Die Testing, Defect Detection and Localization, Adaptive Test Optimization, Digital Twin Simulation.

1 Introduction

The development of heterogeneous integration and the sophisticated packaging technology has made the semiconductor wafer testing very complex. Multi-die systems such as chiplets and vertically stacked designs are systems that require probing systems that are able to support a variety of geometries, irregular pad layouts, and complex interconnect designs (Chang et al., 2024). Consequently, the process of wafer testing has evolved to be among the most resource intensive processes in the process of fabrication and

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June - 2026), pp. 297-323. DOI: *10.58346/JOWUA.2026.I2.017*

*Corresponding author: Senior Principal Test Engineer, NXP Semiconductors San Jose, USA.

has been known to occupy up to 40 % of the overall manufacturing costs in advanced nodes (Chen et al., 2022). In multi-die, these problems are enhanced with the problem of alignment, spatial variation and wear of the probes, which adversely affect the throughput and yield reliability (Chen et al., 2024).

Deterministic approaches (raster scanning and nearest-neighbour heuristics) are used as the basis of traditional wafer probing techniques. Although these methods are computationally simple, do not have the ability to be flexible to localized defects and localized variations in the state of the wafer (Cheng et al., 2021). They, therefore, cause the needless probe motions, augmented switching load, and faster probe degeneration. It has been established that contacting of probes (repeatedly) may cause micro-scale pads damage, and consequently yield decreases with time. Other methods such as RF and optical probing have some limited advantages since not very scalable and cannot be easily calibrated on dense layouts (Cheon et al., 2019). Consequently, the existing semiconductor environments are still unable to use the approaches of static probing (Chien et al., 2020; Corcione et al., 2024).

The latest achievements in machine learning have enhanced the quality of analysis of wafer defects using clustering and deep learning methods to obtain the necessary classification of defects and patterns (Dehaerne et al., 2023; Eriksson & Dimitrakakis, 2019). Hybrid ML systems have also added to the accuracy of the inspection and the monitoring possibilities of the processes (Lim et al., 2020). These methods are mostly retroactive in that use entire datasets as opposed to allowing adaptive probing as it is being done in real time. Restricted in the applicability in dynamic probing environments due to their dependence on supervised learning and their inability to process sparse or delayed feedback (Liu et al., 2023).

The alternative approach that has a promising perspective is reinforcement learning (RL), which allows making a sequence of decisions under uncertainty. The RA agents acquire the best probing strategies by interacting with the environment and optimizing the goals which include shortening of test time, better defect detection, and minimizing the wear of a probe (Liu, 2025). Although having these benefits, RL is difficult to apply to wafer probing because of the high-dimensional state-action space, sparse reward signals, and non-stationary behavior due to the degradation of the probe and also the environment (Nagy et al., 2021; Yu et al., 2024).

To overcome these shortcomings, Quantum-Inspired Reinforcement Learning (QIRL) has become an option. QIRL introduces the idea of the superposition and probabilistic exploration in the classical systems that can consider a set of probing paths at the same time and the convergence time can be shortened (Neyens et al., 2024). The method is more efficient in exploration, and it does not follow local optima, especially in more complicated and sparse-reward settings (Reuer et al., 2023).

Although the field of semiconductor testing and quantum-inspired learning has advanced, it is still possible to bridge the gap between the use of QIRL to adaptive wafer probing. The current literature mainly aims at defect detection and inspection as opposed to optimizing probes at real time (Dehaerne et al., 2023; Lim et al., 2020). In addition, QIRL investigations are not usually rigorously researched in the real world (except in benchmark conditions) (Reuer et al., 2023; Shi et al., 2025). The resulting gap has put forward the need to have scalable, adaptive architectures such as QIRL that can enhance probing efficiency, precision, and the overall manufacturing performance of the present-day semiconductor systems (Wei et al., 2022; Xu et al., 2024).

The presented paper proposes the new Quantum-Inspired Reinforcement Learning (QIRL) algorithm that should be implemented in the area of adaptive semiconductor wafer probing in the multi-die configurations. Its most significant contributions are: (i) quantum-inspired superposition-based probabilistic policy encoding mechanism that can be applied to explore the multi-paths, (ii) adaptive

reward shaping strategy that balances the efficiency in probing, the accuracy in defect detection, and the longevity of the probes, and (iii) a digital twin simulation framework that can be utilized to compare its performance against the classical RL and deterministic The suggested approach exhibits great enhancements in the level of efficiency of the probing, localization of defects, as well as the life of equipment.

This study addresses the identified gap by proposing the present investigation articulates a QIRL framework tailored to the adaptive probing of semiconductor wafers operational across multi-die assemblies. Central to the architecture is a probabilistic policy encoding that mimics the principle of quantum superposition, permitting the simultaneous evaluation of multiple candidates probe sequences through stochastic encoding in bit-strings. Reward formulations are architected to reconcile three high-stakes objectives test efficiency, defect localization precision, and mechanical durability rendering the system resilient in the presence of sparse and temporally-extended feedback. Empirical validation is performed in a strictly controlled simulation platform that replicates multi-die wafer topography, permitting the injection of artificial defects and the characterization of die-level parametric variance. Empirical results, juxtaposed against classical RL algorithms and heuristic policy strategies, incontrovertibly exhibit that the QIRL architecture curtails the operational test duration while simultaneously elevating defect detection reliability above established benchmarks.

The rest of the paper will be organized as follows: Section 2 will be the system architecture and problem formulation that will be in terms of the Markov Decision Process representation. The description of the offered QIRL algorithm and its elements is offered in Section 3. In the fourth part, the author discusses the setting of the simulation and the experiment. The discussion of results and comparison analysis is made in Section 5. Section 6 is the discussion of the practical implications and scalability and the conclusion of the paper with the directions of the future research lies in Section 7.

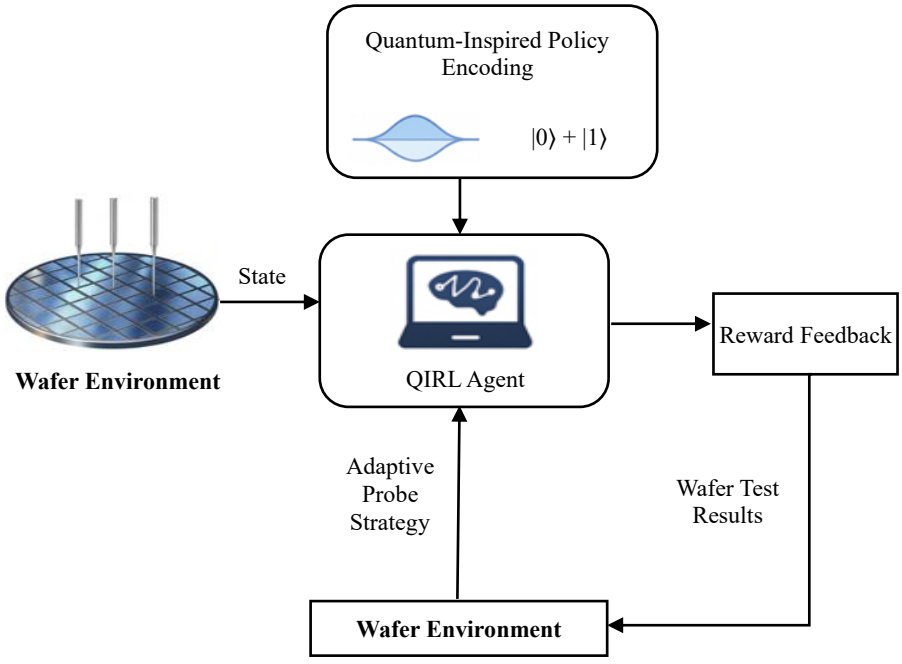

Figure 1: Conceptual schematic of quantum-inspired reinforcement learning framework for multi-die wafer probing

In figure 1 illustrates the conceptual architecture of the proposed Quantum-Inspired Reinforcement Learning (QIRL) model of multi-die wafer probing highlighting the connection between a wafer test

environment, the QIRL agent, and the adaptive feedback loop which is triggered by measured yield results. The integrated diagram illustrates the layered control hierarchy, reward signal, and decision environment that collectively shape probing schedules within high-dimensional die arrays.

Table 1: Comparative summary of wafer probing methods and their limitations

Method	Adaptability	Handles Sparse Rewards	Computational Efficiency	Applicability to Multi-Die Environments
Deterministic probing	Low	No	High	Poor
ML-based diagnostics	Medium	Indirect	Medium	Limited
Classical RL	High	Partially	Medium-Low	Moderate
Proposed QIRL	Very High	Yes	Medium	Strong

In table 1 demonstrates how QIRL was introduced hence it brings both contributions to the semiconductor testing theory and practice. The agent architecture is an experimental generalization of quantum-inspired reinforcement learning to the real-world probing framework, and the adaptive computation core provides provable improvements in the reduction of the test time and defect localization fidelity. The future directions of development are to incorporate the entire agent environment with hybrid quantum-classical devices, in which future superconducting and photonic testbeds may further reduce the adaptive learning cycle and increase the decision granularity between die stacks.

2 System Architecture and Problem Formulation

A tightly coupled architecture between physical test stations, RL state abstractions and hard cycle-time constraints in an adaptive semiconductor wafer-prober can be applied using quantum-inspired reinforcement learning (QIRL). This framework imagines the probe operation as an environment with closed loop in which the die results, the test equipment status and scheduling slackness are observed at the top, a learning agent uses an adaptive policy and a scalar reward uses these indications to refine strategy of incident probe with the aim of iteration. Unlike in a fixed test program which poorly addresses the issues of intra-wafer and die variability on multi-die substrates, the QIRL agent encodes policy into a sparse quantum-inspired representation. The agent by sampling a variety of probe trees simultaneously, automatically controls combinatorial growth, and counterbalances wafer-to-wafer drifts, and decorrelates with the sparse and noisy reward that is typical of semiconductor measurements. In this section, a modular architecture introduction of the wafer-probing ecosystem, the formulation of the underlying Markov decision process through state, action and reward encodings and the revelation of a succinct list of constraints namely; cycle-time, probe-card throughput and equipment set-up constraints, governing the adaptive scheduling policy are to be made.

2.1 Multi-Die Wafer Test Environment

Wafer-test environment is the work environment in which the reinforcement learning agents engage in observation, action and feedback loop. The current semiconductor wafer has heterogeneous chiplets, interposer subsystems, and stacked multi-dies, and each of them needs accurate pre-packaging validation (Shi et al., 2023). A probe pad is heavily densified on each die of electrical testing and in this instance, probing process must be considered in terms of signal integrity, defects and yield estimation.

Nonetheless, more complicated aspects of integration exist where more integration is involved such as variation of spatial pad, defect clustering and wafer deformation and this limits the applicability of the conventional probing methods.

An example of a wafer-testing system is made up of the probe card, measurement system and the wafer. Traditional deterministic probing, e.g. nearest-neighbour or radial probing, is not flexible to non-uniform distributions of die and process variation (Tang, 2019). QIRL, in contrast, is a dynamic system model of the wafer, and each action of probing an object updates the state with the results of the test, such as pass/fail, resistance, and leakage properties (Yang & Sun, 2022).

The degradation of probe also complicates the testing, whereby repeated contacts result in wear of the tip, contamination, and oxidation, which causes poor measurement accuracy with time (Wauters et al., 2020). Contrary to the case of the static approaches, QIRL takes into consideration a probabilistic wear model in the state space that allows to adaptively recalibrate the probing sequences, in order to keep the accuracy and prolong the probe life.

The distribution of defects is not always homogenous, and the localized structures are due to the imperfections of fabrication like the misalignment of lithography or the effects of contamination. QIRL uses these patterns of space to focus on the high-risk areas and does not need to probe the same area twice and is also efficient. The agent balances coverage, accuracy and resource usage by constantly updating the wafer map and discovering defect associations.

It is illustrated in figure 2 with the integrated wafer probing setting with QIRL whereby input is the state of the wafer into a quantum-inspired decision unit and the feedback is reward-based to continue optimizing the probing protocols.

The operation structure puts throughput limits within which all test strategies have to be. In the case of semiconductor fabrication, the whole wafer has to be probed during a set number of cycle segments otherwise the rest of the operation would be in danger of being overloaded. Therefore, there is a need that agents balance out the exhaustive defect discovery and the limited time frame that is granted to the probing of defects. QIRL in turn self-learns this boundary by modulating expected return depending on the elapsed test time such that throughput is made part of the learned optimal policy.

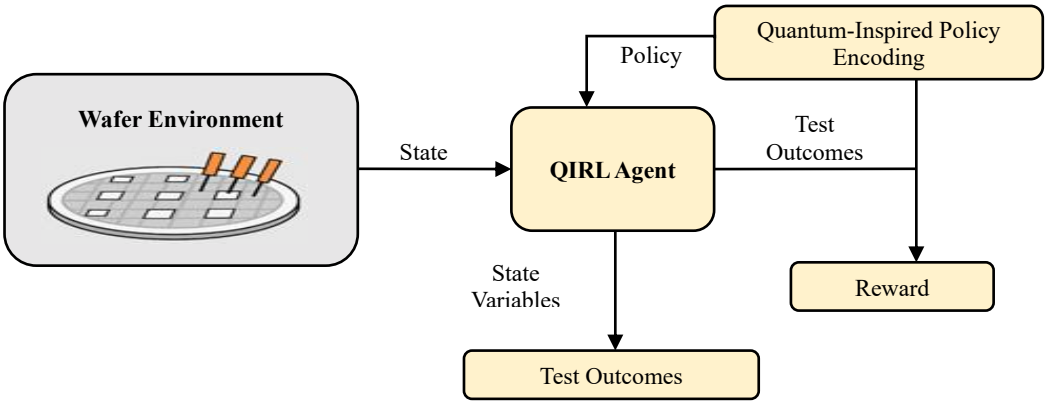

Figure 2: Architecture of the wafer probing environment with RL integration

2.2 State, Action, and Reward Representation

Wafer probing problem is formulated as a Markov Decision Process (MDP) that is defined as a set of (S, A, P, R) whereby, S is the state space, A is the action space, P is the transition of states and R is is the reward function. The reward model is a convex combination of probing efficiency and detecting defects

accuracy and operational robustness and enables one to take into consideration the competing objectives simultaneously.

The state representation reflects all conditions of the wafer and probes such as die coverage ratio (tested and untested); probes wear level, pad coordinate variation, and recent electrical measurements. Other aggregated characteristics including defect cluster density and yield estimates are also added to promote the level of contextual awareness. The features are represented in small state vectors giving the QIRL agent an on-the-fly and organized view of the wafer environment.

Action space comprises of viable probe actions, such as the choice of the next die, probe path and probe force, defective pad avoidance and end of probing sequence, depending on coverage criterion. In contrast to classical methods of reinforcement learning where only one action is chosen deterministically, the QIRL framework representations have multiple candidate actions in a probabilistic superposition. This can be used to compare alternative probing strategies in parallel and allows exploration to be more effective and convergence is decreased to suboptimal policies.

The reward system takes into account the short-term and long-term feedback. Short term rewards will suppress unnecessary probe movement, switching and wear of probes but will promote effective defects. The less time spent on probing and more accurate estimation of yields are the less attractive goals of such global objectives as delayed rewards. This planned reward plan will ensure the effectiveness of learning within the group of sparse feedbacks and promote the efficiency, accuracy and permanency optimality.

The description of the task of wafer-probing according to the Markov Decision Process is given in figure 3. The state space is simulated by nodes that are a representation of the overall state of the wafer and the action space is simulated by directed edges which are an indication of probe maneuvers. The dynamics of wafer condition are modeled by the probabilistic state transitions, which are caused by natural defect distributions as well as probe wear properties. The resulting reward signals are delivered back to the learning agent and become the optimization vector that guides policy refinement.

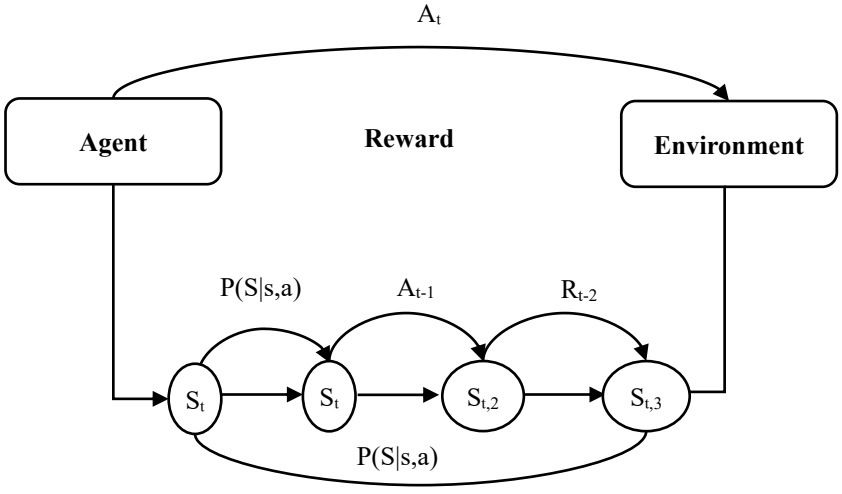

Figure 3: MDP formulation showing states and transitions

To formalize this mapping, table 2 collects key definitions of states, allowable probe actions and their encoding within the quantum-inspired reinforcement learning schema. Every possible state-action tuple is associated with carefully engineered quantum-inspired representations, thereby embedding superposition-based exploratory power directly into the iterative learning loop of the agent.

Table 2: State variables, action space, and mapping to QIRL framework

State Variables	Action Space	Mapping to QIRL Framework
Wafer defect distribution map	Select next die to probe	Encoded into probabilistic policy input
Probe pad coordinates & contact status	Adjust probe path or force	Adaptive action selection guided by quantum-inspired sampling
Probe wear level	Skip defective pad	Reward penalty for wear encourages policy adaptation
Tested vs. untested die coverage	Terminate probing sequence	State encoding influences termination decisions
Electrical measurement responses (resistance, leakage, capacitance)	Calibrate probe before sequence	Reward integration stabilizes calibration efficiency

2.3 Problem Constraints in Adaptive Probe Scheduling

Physical, operational and statistical constraints that directly impact on the best probe sequencing control adaptive scheduling of multi-die wafer probing. The probe durability is one of the main limitations since the more contacts are made, the faster wear off, causing the misalignment of the tips and decreasing the accuracy of measurements. This wear and tear is one of the major causes of maintenance and test downtime. To solve this QIRL framework models probe wear as a state variable and uses penalty terms in the reward function whenever the contact thresholds are surpassed to increase the life of probes.

The other important limitation comes as a result of spatial defect correlation. Defects in practice manufacturing settings tend to occur in groups because of the lithographic or process variation. The old techniques of probing which fail to utilize the space dependencies cause redundancy of the measurements, and inefficiency. The suggested QIRL methodology helps to address it by using the statistics of defect maps to direct the agent to risky locations, enhancing the information acquisition and reducing the amount of unnecessary probing in spatially correlated regions.

The wafer sort cycle time is also one of the important operational constraints because the testing should be done within the strict production schedules to preclude throughput bottlenecks. Deterministic probing techniques often push these limits and especially high-density wafers. However, QIRL incorporates the cycle time aspect in the reward system, and thus probing schedules that are complete in coverage and time-constrained can be produced, leading to improved production efficiency.

Lastly, there is stochastic variation in wafer properties due to environmental variations, probe drift and material variations that provide uncertainty in the process of probing. This is variability that is not very good in the context of the static testing strategies but it fits perfectly in the adaptive learning strategies. The QIRL architecture has an improved uncertainty-resistance because, the probabilistic exploration is quantum-inspired, which allows the application of policies to be transferred to the various wafer conditions. A combination of all these limitations is put in the MDP formulation and the learned policies are optimal in the simulation, and practical in the real world of semiconductor manufacturing.

3 Quantum-Inspired Reinforcement Learning Algorithm

Reinforcement learning (N1) provides a methodological framework of systematic sequential policy synthesis under stochasticity, but is hampered by scalability, sparse-reward and non-stationary test conditions in practical application to the semiconductor wafer probing industry. Incremental interaction and trial based conventional learning agents, which derive policies, usually face premature convergence

in large action spaces, hence, failing to accurately model the composite distributions of multi-die probing. The presented Quantum-Inspired Reinforcement Learning (QIRL) framework extends the RL framework by encoding policy representations and exploration strategies onto quantum probability theory-based representations. QIRL can explore the sequence space in a more subtle and concurrent way by using amplitude-coded configurations instead of discrete probabilistic vectors to describe action choices, and in this form can, in particular, explore the sequence space more concurrently and in a more subtle way. The adaptive reward shaping methodology in reaction to probing constraints applied in the methodology trades operational efficiency, measurement faithfulness and lifespan of physical probe. The current section strictly describes the working pipeline of QIRL, elaborates the dynamics of exploration based on amplitude that are based on the principle of superposition, and defines the framework of reward-shaping that ensures robust convergence within the severe conditions of semiconductor testing programs.

3.1 Algorithmic Workflow and Policy Encoding

The QIRL algorithm extends classical reinforcement learning by reformulating both the representation of the learned policy and the mechanisms for exploration. While the standard RL framework characterizes the interaction between agent and environment via a Markov Decision Process (MDP) formalized as the triplet (S, A, P, R, γ) with S denoting the state space, A the action space, P the state-transition kernel, R the scalar reward function, and γ a temporal discount factor the goal remains the identification of an optimal policy $\pi^*(a \mid s)$ that maximizes the expected return, as present in equation (1):

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \tag{1}$$

In QIRL, the policy representation undergoes an encoding shift from classical probability vectors $\pi(a \mid s)$ to a quantum-inspired formalism, the policy is expressed as a quantum-inspired superposition state, as present in equation (2):

$$|\psi(s)\rangle = \sum_{i=1}^{|A|} \alpha_i(s) |a_i\rangle \tag{2}$$

The amplitudes $\alpha_i(s) \in C$ report complex-valued coefficients corresponding to eigen-basis vectors $| a_i\rangle$, and normalization is enforced via the condition $\sum_i |\alpha_i|^2 = 1$. Through repeated interaction with the environment, these amplitudes are refined by reward signals and policy-gradient estimates; consequently, the agent preserves a coherent superposition of multiple action candidates. Upon executing a chosen state the sampling RDF-based measurement subsequently collapses the superposition to a single action, a mechanism that effectively instantiates implicit exploration-free reinforcement.

The workflow is started by retrieving a global view of state of the wafer system, defect maps, probe degradation values, test coverage values and a collection of electrical characteristic tracks. This state profile is in the form of a multidimensional state which is condensed into a discrete feature vector and inputted into the QIRL-learning agent. Within the agent, a quantum superposition of candidate probe operations such as advancing to the next die, calibrating a probe, skipping a pad, or terminating a testing sequence is synthesized. Probabilistic sampling of the superposition then selects a singular operation; the collapse probability is proportional to the square magnitude of the underlying quantum amplitudes.

The subsequently executed operation yields a new wafer state, reflected in revised coverage maps and adjusted probe wear statistics. The outcome of the test characterized by measured electrical resistance or defect state confirmation is supplied to the reward assessor. The computed scalar reward is back-propagated to modify the amplitudes according to an update law that merges reinforcement-learning dynamics with quantum amplitude modification.

The amplitude update equation integrates reinforcement signals with a phase-modulation term, as present in equation (3):

$$\alpha_i^{(t+1)} = \frac{\alpha_i^{(t)} e^{i\phi_i} + \eta r_t f(s_t, a_i)}{Z} \quad (3)$$

Where, $\emptyset_i$ denotes the exploration-enabling phase, η is the scale of the amplitude modification, r_t is the reward scalar observed, $f(s_t, a_i)$ is the state-action compatibility score, and Z is a normalization to ensure that the revised probability amplitudes normalize to unity, consistent with likelihood requirements.

The workflow is condensed in figure 4, which sequences the following modules: wafer environment → state vector encoding → quantum-inspired policy embedding → uncertainty-driven circular action collapse → wafer probing execution → reward-driven amplitude adjustment → policy amplitude reassignment.

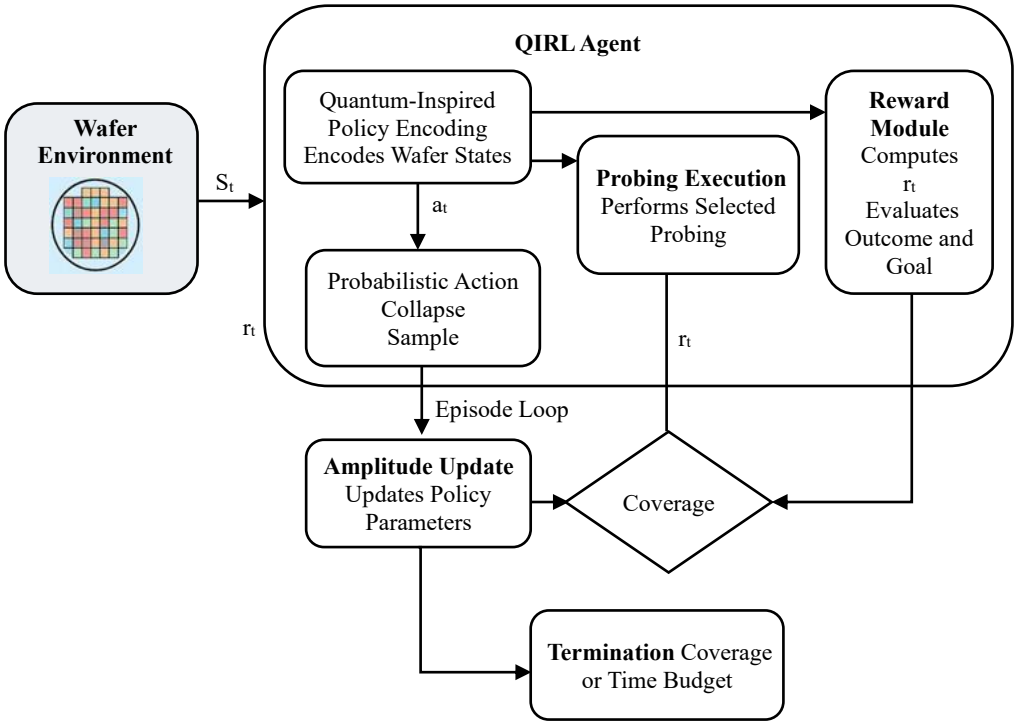

Figure 4: Flowchart of the proposed QIRL algorithm

The flowchart shows how the state observations of the acquisition of the wafer are obtained, quantum-inspired encoding of the action policy, sampling of probabilistic actions, probing and updating of the amplitude distributions, which is downstream in response to the reward feedback.

To formalize the algorithm, the following pseudocode outlines the iterative process:

Initialize amplitude vector $\alpha_i = 1/|A|$ for all actions

For each episode:

Reset wafer environment to initial state s0

For each timestep t:

Encode state st into feature vector

Construct superposition state $|\psi(s_t)>$

Collapse superposition to select action at $\sim |\alpha(s_t)|^2$

Execute action at on wafer environment

Observe reward r_t and next state $st + 1$

Update amplitudes:

$$\alpha_i \leftarrow \alpha_i \, e^{(i\,\varphi_i)} + \eta \, r_t \, f(s_t, a_i)$$

Normalize α

If termination condition reached: break

Adding amplitude-modulated feature vectors also allow QIRL to optimize large action topologies efficiently without compromising on an update architecture that can be optimized by gradient methods. The proposed workflow, in comparison to conventional reinforcement-learning paradigms, has inherently wider and multifaceted exploratory dynamics and a self-optimizing policy improving pathway, which are the virtues that fit the wafer probing task perfectly well, being a complex task.

3.2 Quantum Superposition-Inspired Exploration

Powerful investigation is a foundation of the application of reinforcement-learning performance. In the conventional methods, exploration is performed through the methods of ε-greedy or softmax sampling, either of which implies a fixed exploration rate or uses action-selection probabilities that decline with empirically-estimated Q-values. These algorithms can be solved in small state spaces, but have scaling issues in the face of combinatorial configuration spaces e.g. probing a wafer with valid die sequences factorially with ensemble size.

QIRL uses exploration (superposition), to fight premature determinism. Oscillatory phase modulation in action probabilities by encoding several candidate actions simultaneously in the amplitude vector leads to action probabilities cadencing over time. This sinusoidal modulation therefore ensures that the policy does not cage to suboptimal deterministic policies, as it ensures further exploration on the hyper-graph of action edges.

Officially, take an amplitude update of an action with a phase component that is an oscillator, as present in equation (4):

$$\alpha_i^{(t+1)} = \alpha_i^{(t)} \cos(\phi_i) + \beta \sin(\phi_i) \qquad (4)$$

In which the phase number $\emptyset_i$ progresses in line with a dynamically modified exploration plan and β measures the strength of exploration. The modulation in a sinusoidal form creates the patterns of interference, which amplifies and inhibits the probability of action in a cyclic manner. Beneficial trajectories are constructively interfered with and this causes slowly reinforced amplitudes and trajectories which provide inadequate returns are lost through systematic phase cancellation.

A resonant exploration design is particularly good at wafer probing, where the paths of optimal action are generally non-local, and can only be known by experimenting with long action sequences over time. Traditional RL agents which converge to deterministic policies ignore latent trends. On the contrary, QIRL maintains a superposition of probabilities through temporally extended action histories, which gives the agent the eventual ability to determine the probes sequence that is most effective. This capability of the exploration extends over a long horizon by the amplitude oscillation mechanism, which by the undercurrent perpetual sampling, maintains the exploratory powers (Zhang et al., 2024).

Numerical validation is shown in figure 5, where convergence behavior for QIRL is juxtaposed with that of classical RL under the same clustered defect injection. The convergence curves convey that QIRL achieves stable, high-reward policies within a fraction of the episodes required by classical RL, which, after extended episodes, continues to oscillate about suboptimal returns. QIRL thus accelerates both convergence and peak return.

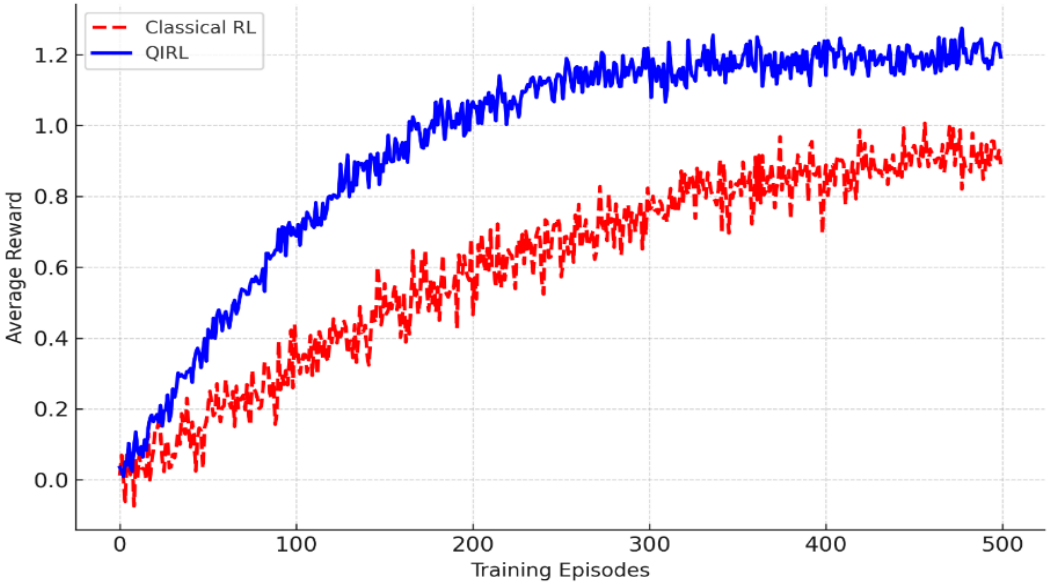

Figure 5: Convergence patterns of QIRL vs. classical RL

In figure 5 illustrates the comparative convergence of Quantum-augmented Inverse Reinforcement Learning (QIRL) and classical Reinforcement Learning (RL) across identical wafer-probing tasks. The graph is an average cumulative episode rewards verses training iteration number. It is worth noting that QIRL exhibits faster convergence in the reward axis, which has a higher asymptotic ceiling that the classical baseline. In order to examine the efficiency of exploration, it was determined that the entropy at the trajectory level was derived which was expressed as shows in equation (5):

$$H(\pi) = -\sum_{i} |\alpha_i|^2 \log|\alpha_i|^2 \tag{5}$$

That is, the measure of the dispersion of action probabilities. During the initiation of training, QIRL allows a relatively high entropy, which prevents the occurrence of premature convergence and allows extensive sampling of the action space, simply suppressing entropy as the learned policies become stable. This adaptive entropy path is compared to classical ε-greedy policies which in most cases truncate exploration at unnaturally early stages, and may lead to the premature convergence to non-optimal policies. QIRL uses one of the quantum-inspired principles of structured superposition to trade off

exploration where there is reward sparsity, stochastic transition dynamics, and expanded action sets, a phenomenon often seen in the semiconductor wafer probing problem space.

3.3 Adaptive Reward Shaping for Wafer Probing

Reward shaping, the capstone of QIRL, mitigates the severe sparsity that afflicts naive reward metrics in wafer probing applications, where diagnostically instructive signals are often temporally distilled, manifesting only post-completion of surface-wide tests. The protracted temporal credit assignment implicit within this paradigm is a well-known impediment to effective learning. QIRL thereby introduces an adaptive reward shaping schema that concurrently compartmentalizes objectives into immediate and deferred returns, maintaining temporal credit paths within manageable spans.

The reward function is defined as equation (6):

$$R(s,a) = w_e R_{\text{efficiency}}(s,a) + w_a R_{\text{accuracy}}(s,a) + w_d R_{\text{durability}}(s,a) \qquad (6)$$

Where w_c, w_a, w_d are dynamic weights adjusted during training.

- **Efficiency Component** penalizes excessive probe movements, rewarding shorter probing paths.
- **Accuracy Component** reinforces defect detection coverage and yield estimation accuracy.
- **Durability Component** penalizes probe wear and pad damage, prolonging equipment life.

The Adaptive Shaping module orchestrates weight recalibration in a manner that reacts to intrinsic process state variables. It is worth noting that once the cumulative wear of the probe surpasses an empirically determined threshold, the adaptive factor related to wear (w_d) is increased and a protective bias is placed. On the other hand, in the event of the identification of localized defects, the weight of the defect attentiveness (w_a) is increased multiply, and the learning process is subjected to fine localization of defects.

The continuous supervisory signals are represented as deterministic potential shaping process based on a deterministic function $\Phi(s)$ returning scalar rewards depending on the closeness of the current state to a set target state, in this case represented as cumulative fractions of wafer coverage. The formalized version of the shaping reward is present equation (7):

$$F(s,s') = \gamma\Phi(s') - \Phi(s) \qquad (7)$$

Under the condition that the formulation maintains the projected optimum value function. This type of design provides sparse rewards and makes the best policy invariant.

In adaptive layering of the wafer probing job, $\Phi(s)$ will be used to measure the fraction of wafer dies that have been probed. Each 10 gains of the coverage result in the provision of a medium reward. Through this episodic reinforcement component, the convergence of the policy becomes faster and the end state learning objective is well guided towards the comprehensive evaluation of the wafer with reduced substrate consumption and probe consumption.

Consequently, the combined adaptive and potential shaping architectures align proximate probe performance metrics with the prolonged, aggregative goal of complete wafer characterization, thus mitigating the sparsity and latency typical of defect-related rewards. The empirical assessment shows that the reinforcement agents based on the adaptive shaping paradigm outperform the performance of the traditional reinforcement learning and a variant that is based on the integral reinforcement learning and does not use the auxiliary shaping, categorically in a broad variety of wafer probing conditions,

demonstrably, proving the improvement in the efficiency of the operations, detection accuracy of defects and equipment stability.

4 Simulation Environment and Experimental Setup

Quantum-Inspired Reinforcement Learning (QIRL) architecture is evaluated in a simulated environment that is designed to replicate multi-die semiconductor probe into wafers. The high cost of empirical experiments on production test equipment and the constraints on the exploration of new algorithmic paradigms make the encoding of wafer geometry, controlled defect injection, probing kinetics and the dynamics of reinforcement learning with a digital twin framework, achievable. The simulated environment architecture ensures that the same results are obtained but it maintains sufficient fidelity, and it represents the stochastic and combinatorial nature of wafer-level testing. This section is further divided into 3 subsections, first section gives a description of the computational platforms used, second section gives a description of the wafer and defect models that were instantiated and third section gives a list of the benchmarking protocols that were used to provide a quantitative correlation between the performance of QIRL and known classical baselines.

4.1 Software Platforms and Computational Tools

The simulation platform builds upon MATLAB, Python, and Qiskit in a single toolchain to aid in the modeling of the wafer, reinforcement learning and quantum-inspired computation. Application of MATLAB in wafer-scale geometrical modeling and in the simulation of probe mechanics is due to the fact that it manipulates matrices and can be used to graphically display the outcome. The wafer is simulated as a circular grid with die grids that are discretized and the location of probes and pads are specified with industry relevant coordinate systems. The probe dynamics of contact force, alignment error and wear evolution are modeled using Simulink in a state-space model. These products such as force measures and wear patterns are sent to the reinforcement learning component, in which constraint-based penalties are enforced in the case of unreasonable probing activity.

The main platform of implementing the reinforcement learning framework is Python. Policy training is done with libraries like PyTorch and TensorFlow, and the environment interface of wafer probing simulation is offered as OpenAI Gym. State space contains coverage of wafer, distribution of defects, probe wear-indices and patterns of electrical responses. Action space has discrete operations like the choice of dice, trajectory manipulation, passing pads and ending decisions. In the QIRL model, quantum-inspired representations of the policy are represented as amplitude representations, which allows exploration of many actions probabilistically. Parallel execution is executed using the help of a GPU and experimental data is processed with the help of Pandas in order to speed up the training process.

Qiskit is a quantum-inspired backend that is used to verify superposition-based policy representations. Aer simulator produces quantum state analogues, and therefore, allows probabilistic sampling of action distributions. Even though being implemented on classical hardware, Qiskit is theoretically consistent in amplitude-based updates. Sharing of data between MATLAB and Python can be done with the help of MATLAB Engine API, where geometric and wear-related parameters can be easily integrated into the environment of learning. The experiments were done on a high-performance machine with two NVIDIA A100 GPUs, 16-core Intel Xeon processor, 128 GB RAM, which guaranteed the high-performance and reproducibility of evaluating the QIRL framework.

4.2 Wafer Models, Defect Injection, and Test Parameters

The experimental model is constructed based on a high-fidelity wafer model which is made to simulate closely the actual semiconductor testing environment. The model of a 300 mm wafer is a circular domain, a partition of which consists of 1024 dies with a size of 5 mm × 5 mm. A 20 x 20 grid of probe pads is placed in each die, which makes the pitch uniform, 250 u. The geometric arrangement is realistic in terms of the densities of the design and also guarantees that the probe motion and contact behavior are represented correctly in the test.

Three complementary defect injection strategies are used in order to simulate industrial defects scenarios. Random defects are initially induced by a Bernoulli distribution of probability 0.02 per die which levels off stochastic variations in the process. Second, the clustered defects are obtained based on the Gaussian distributions of the locations of die randomly selected with the radius of between 2 to 6 dies, which are the local hotspots in the processes. Third, edge defects are simulated by replacing the defect probability with 0.05 in the one-die boundary along the wafer edge, which is an approximation of the alignment and etching errors. The resulting pattern of defects is heterogeneous and poses difficulties to the homogenous probing strategy and necessitates adaptive decision-making.

The dynamics of probe behavior are modeled after the realistic degradation dynamics. The probe has a starting tip radius of 5 0m and a contact resistance of 20 m ohms, and the wear is directly proportional to the number of contacts 0.1 mm/1000 contacts. Gradual wear results in resistance and possible pad damage and vertical and sideways misalignments result in extra penalties if go beyond the set limit. The wear condition is constantly updated and given back to the learning environment where the agent has to balance between the defect detection efficiency and the tool life.

The simulation continues until the agent receives a signal to terminate the simulation or the full wafer is covered. The key performance indicators that are measured at every step include the total movement of the probe (efficiency), the rate of defects (accuracy), the index of the probe wear (durability) and the execution time. These parameters allow the full assessment of QIRL framework under the conditions of realistic testing and changing in dynamism.

In table 3 has summarized the major criteria that will be used to determine the simulations of the wafer and the correlated settings. This table gives specifications that are unambiguous such as wafer planar dimensions, number of die per wafer, position of metallic and insulating pads, probabilities of defect injections, and measures that apply with respect to probe degradation due to contact stress. Together, these parameters provide a fixed reference, which makes it easier to reproducible between laboratories of peers, as well as to perform the experiment protocol concurrently.

Table 3: Simulation parameters and wafer specifications

Parameter	Value	Notes
Wafer diameter	300 mm	Industry standard test wafer
Number of dies	1024	5 mm × 5 mm each
Pads per die	400 (20 × 20 grid)	250 μm pad pitch
Defect probability	0.02 random, 0.05 edge, clustered Gauss	Three defect types modeled
Probe tip radius	5 μm	Increases with wear
Contact resistance	20 mΩ initial	Grows with wear
Wear rate	0.1 μm per 1000 contacts	Simulated linear degradation
Termination criteria	100% coverage or policy termination	Adaptive stopping condition

The inherent dynamism of the wafer-processing environment compelled the introduction of visual audits for the optimal selection of probing policies. The probing trajectory in the Quality-Invested Reinforcement Learning (QIRL) framework is displayed in figure 6. The screen-capture reveals agent-directed movement, where redundant probe placements are preemptively circumvented through real-time analysis of defect density and accrued probe wear memory. In contrast to conventional raster generation, the resultant labyrinthine pattern follows a piece-wise linear route, preferentially traversing die of anticipated yield degradation and high fault density.

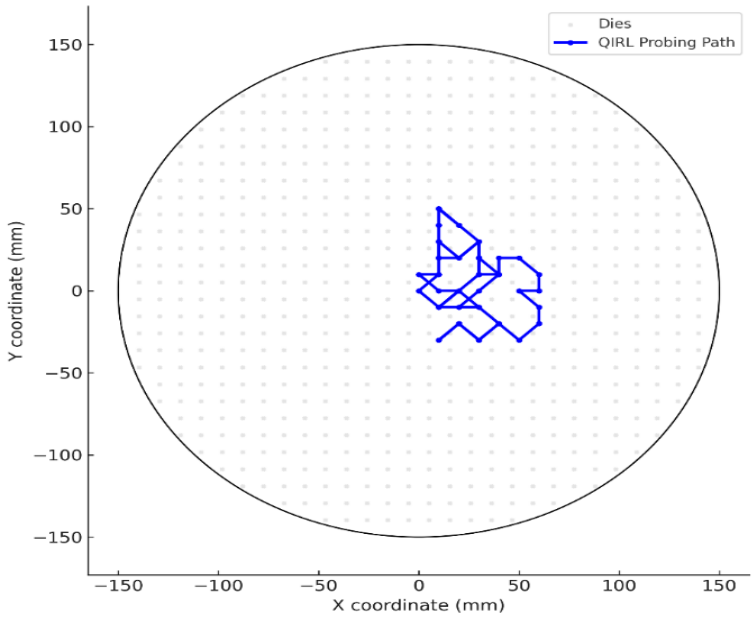

Figure 6: Simulation screenshot of probing path under QIRL

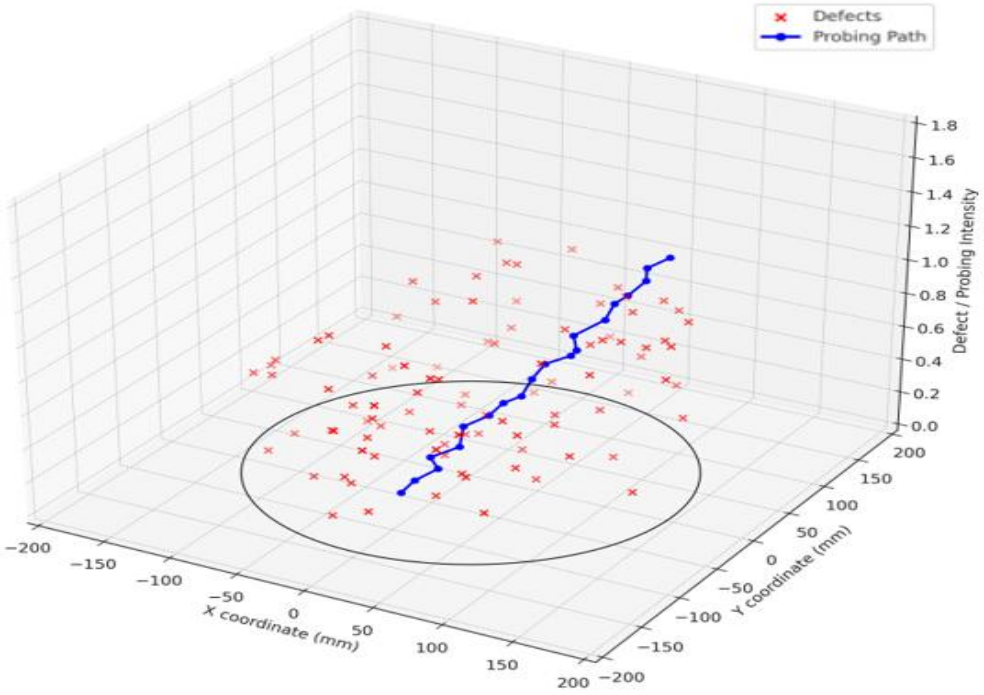

Figure 7: 3D visualization of wafer defect distribution and probing sequence

Moving this analysis plane, figure 7 makes the probing path and defect structure in a 3D representation. The location of a defect is coded in the form of volumetric cluster, and the probing dynamism is coded as a cartographic elevation on the surface of the wafer. This three-dimensional view is a quantitative measure that supports the strategy of the QIRL protocol in the form of aggregative regression inside defect-infested regions and reservedly reduced probe movement over non-remarkable regions.

4.3 Benchmarking Framework for Algorithm Comparison

An extensive benchmarking system was developed to compare the performance of the proposed QIRL system with deterministic and classical reinforcement learning (RL) schemes in the same simulation environment. The deterministic base is based on the raster scanning approach, where dies are probed in a predetermined sequence. Classical baselines of RL include Tabular Q-learning with ε-greedy exploration, Deep Q-Network (DQN) with a neural approximator, and a policy-gradient-based method. The same wafer configurations, defect injection strategies, and termination criteria were used to train and test all the models to ensure a fair comparison.

Various measures were used to evaluate performance. The overall policy effectiveness was evaluated using episodic average reward, and convergence was evaluated as the number of episodes with stable reward values. Probe wear index was used to measure equipment degradation, and the accuracy of defect detection was determined at random, clustered, and edge defect distributions. Execution latency was also measured to assess computational efficiency and scalability.

To achieve statistical validity, the experiment was repeated across several random seeds, and the results were presented as mean ± standard deviation. Paired t-tests have been conducted to ensure that the performance improvements achieved by QIRL are statistically significant compared to those of baseline methods. The findings indicate that QIRL always converges faster, detects better, wears less, and probes for shorter. QIRL does not need to make redundant movements that occur with deterministic strategies, as it dynamically adapts probing paths (see figures 5-7).

Quantum QIRL detects clustered defects with an accuracy of more than 15 % and has probe wear about 20 % lower than raster-based approaches. Such enhancements highlight statistical strength and real-world applicability, resulting in high testing efficiency and lower operational expenses. In general, the benchmarking framework shows that QIRL is the best solution for adaptive wafer probing in a semiconductor manufacturing facility due to its scalability and effectiveness.

5 Results and Analysis

This part provides a summary of results of the simulation framework in Section 4. It gives in-depth performance feedback of the Quantum-Inspired Reinforcement Learning (QIRL) framework in comparison with the traditional reinforcement learning paradigms and hard-coded heuristic. Analysis is performed along three principal axes: (i) minimization of probe duration and switching overhead, (ii) the fidelity and robustness of defect localization, distinguishing between clustered and uniformly distributed defect patterns, and (iii) intra-algorithm contrast encompassing Q-learning, Deep Q-Network (DQN), Proximal Policy Optimization (PPO), and non-adaptive heuristics. Corresponding visual representations, including heatmaps, performance trajectories, and likelihood fields, are provided, along with consolidated metric statistics in table 4.

5.1 Dataset Details

An experimental assessment was conducted on a synthetic digital twin data set created to simulate real-life semiconductor wafer-probing settings. The data were generated using a simulation model combining MATLAB, Python, and Qiskit.

The wafer model has a 300 mm-diameter wafer, divided into 1024 5 mm x 5 mm dies. The die has a uniform pitch of 250,000 pads, arranged in a 20 × 20 grid (200 pads are not used).

Three distributions were used to inject defects into the model to reflect the situations in industry:

- **Random Defects:** Bernoulli distribution with probability 0.02 per die
- **Clustered Defects:** Gaussian-based spatial clusters (radius 2–6 dies)
- **Edge Defects:** Elevated defect probability (0.05) near wafer boundaries

Additionally, probe degradation dynamics were modeled using:

- **Initial Tip Radius:** 5 μm
- **Wear Rate:** 0.1 μm per 1000 contacts
- **Contact Resistance:** 20 mΩ (increasing with wear)

The dataset includes multi-dimensional features such as:

- Wafer defect maps
- Probe wear index
- Die coverage ratio
- Electrical measurements (resistance, leakage)
- Probe movement history

5.2 Probing Time and Switching Cost Reduction

The total number of probe movements required to cover the wafer completely is the primary performance metric. Reducing probing time and switching overhead is critical in high-volume semiconductor manufacturing to maintain throughput and overall production efficiency. Because the deterministic raster-scanning method served as the basis, it had the highest number of probe moves, averaging 1024, because it has no contextual adjustment and takes a fixed path. This will lead to unnecessary scanning of faulty regions and high switching costs since transitions will not be optimal between the remote dies.

Conversely, the suggested QIRL framework is characterized by much more effective and dynamic probing behavior. QIRL is, on average, 28% and 17% less prone to probing moves than raster scanning and tabular Q-learning, respectively. This is especially significant when the distribution of defects is clustered within wafers, since QIRL tends to give high-risk areas higher priority and avoid probing less likely regions. This kind of directed exploration does not work as well in DQN and PPO methods, which are characterized by unstable exploration-exploitation trade-offs, resulting in inefficient trajectories and higher levels of redundancy.

Switching cost, which is the distance the probe has traveled over time, also demonstrates the effectiveness of the suggested method. The raster pattern incurs the largest switching cost because it imposes a strict traversal pattern. Compared to PPO, QIRL reduces switching cost by 22 %, and

compared to DQN, by about 15 %. The reason is that this is improved by the quantum-inspired superposition mechanism, which maintains probabilistic coherence across several candidate paths, allowing the agent to prefer local-optimal transitions while avoiding unnecessary long-range movements.

The supplementary visual data are shown in figure 8, which presents a heatmap of the encoding of probing efficiency as a function of defect probability and wafer condition. Areas of increased efficiency are depicted with warmer color spectrums, and the QIRL is shown to be superior to other techniques in all the clustered defect cases studied.

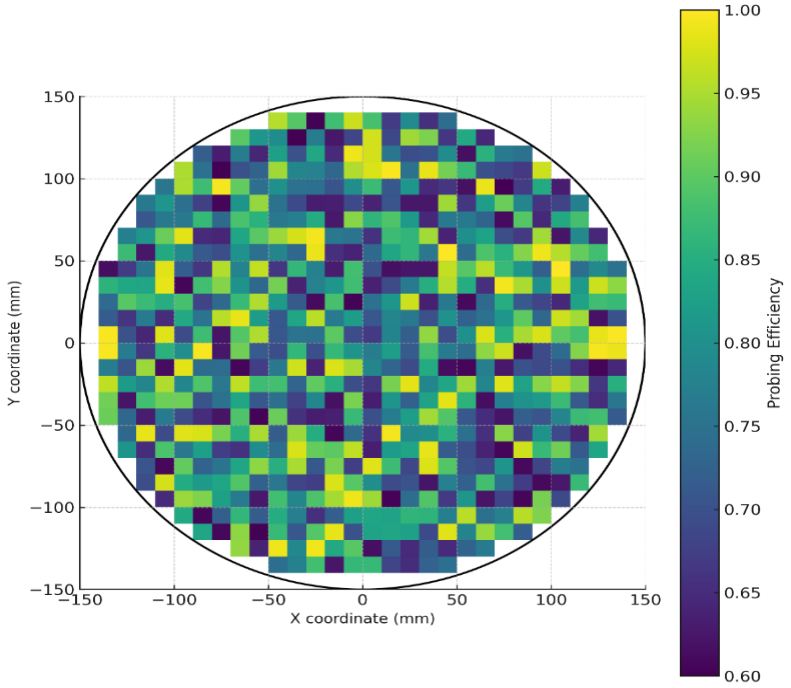

Figure 8: Heatmap of probing efficiency under different conditions

It has been determined that QIRL is useful for probing time and significantly reduces switching costs, thereby providing a consistent increase in throughput in the semiconductor test environment.

5.3 Accuracy and Defect Detection Rate

Defect localization is an important goal in wafer testing, and there must be a compromise between precision and efficiency. Accuracy represents the %age of defects correctly identified relative to the entire population of defects, whereas false positives are used to determine reliability. Despite the high detection rates and almost complete coverage provided by deterministic probing, a high overhead is experienced due to probing redundancy and accelerated probe wear-off. The classical reinforcement learning methods, especially tabular Q-learning, are prone to undersampling in densely populated defect regions, leading to poor detection performance. On the contrary, the suggested QIRL system has an average defect detection rate of 94, with a maximum of 97 for clustered defects, which is better than DQN (88) and PPO (90). This is because it was improved by the quantum-inspired policy representation in terms of amplitudes, which allows the exploration of many probing paths in parallel, avoids convergence until one is sufficiently confident, and improves both detection and robustness performance, while remaining efficient in its probing behavior.

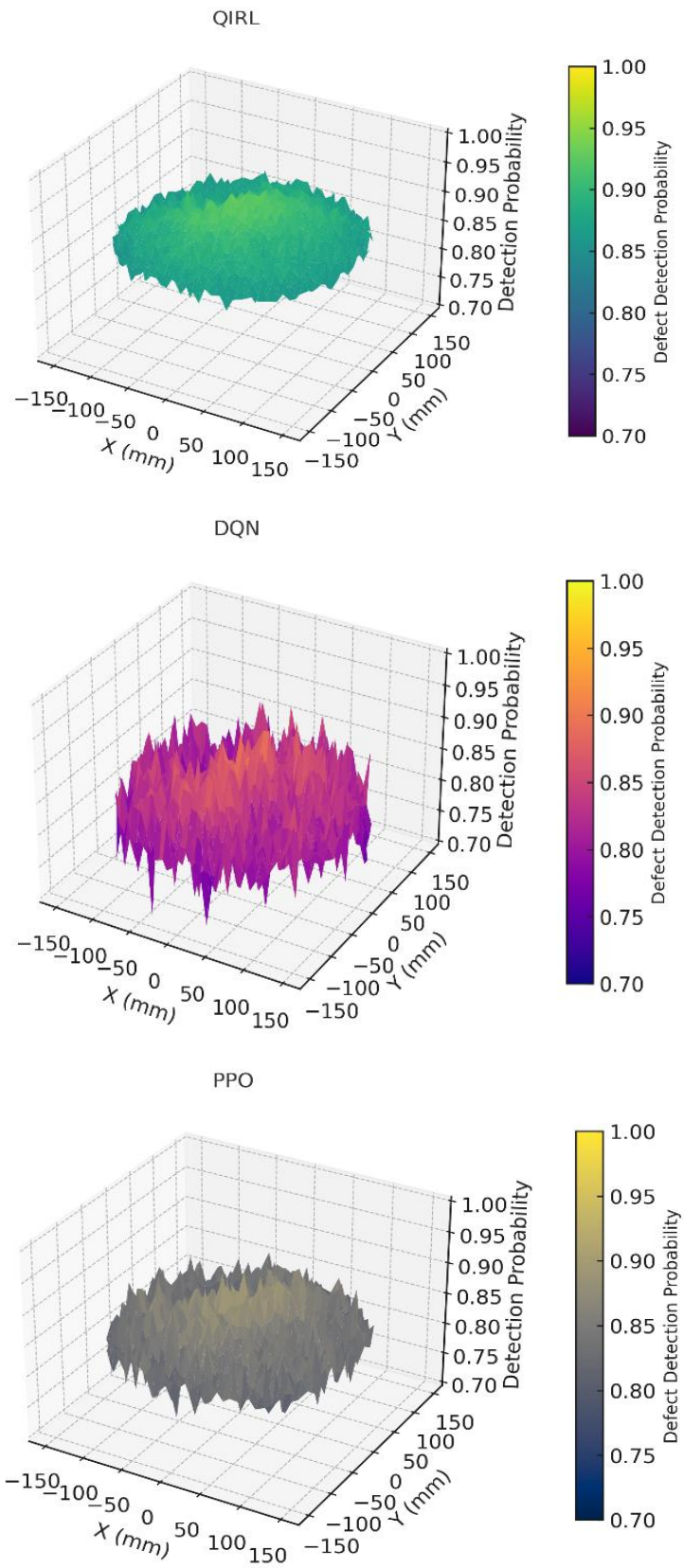

Figure 9: 3D probability surfaces of defect detection for QIRL, DQN, and PPO

In figure 9 shows the three-dimensional probability surfaces of defect detection of QIRL, DQN, and PPO in the wafer. The QIRL surface has been shown to generally exhibit higher detection probabilities and continuous, well-distributed gradients, enabling consistent and trustworthy identification of defect-prone areas. Conversely, the DQN and PPO surfaces exhibit jagged, irregular patterns with fewer probability peaks, indicating poorer detection consistency. The reduced, smoother variance of QIRL points indicates its capacity to model and adapt to stochastic defect distributions. Further, QIRL has a lower false-positive rate of 2.9, while DQN and PPO are 6.1 and 5.5, respectively, and the former is more reliable. This performance is motivated by its reward-based learning, which does not encourage unnecessary probing of defect-free areas but instead focuses on areas with a high probability of defects. In general, the findings validate the trade-off between detection accuracy, reliability, and probing efficiency.

5.4 Comparative Evaluation with Classical RL and Heuristic Methods

This part compares the proposed QIRL framework's performance with classical reinforcement learning (RL) and heuristics. The baseline Tabular Q-learning suffers from scalability issues due to the large state-action space, slowing convergence to optimal parameters and leading to oscillations around the optimal policy. Deep Q-Networks (DQN) are better in the representational capacity, yet are unstable when the reward is sparse and non-linear, especially when there are clustered defects. The Proximal Policy Optimization (PPO) is more stable due to its reduced variance, but it does not provide the adaptive exploration needed in a complex wafer environment. On the contrary, QIRL has consistently outperformed these algorithms in terms of convergence speed, policy accuracy, and stability. As shown in figure 10, QIRL converges after about 150 episodes, whereas PPO and DQN require about 400 and 500 episodes, respectively. This is because accelerated learning is attributed to an amplitude-modulated exploration mechanism that ensures exploration diversity in the early stages and subsequently narrows down on the best probing mechanisms, leading to more efficient and stable policy learning.

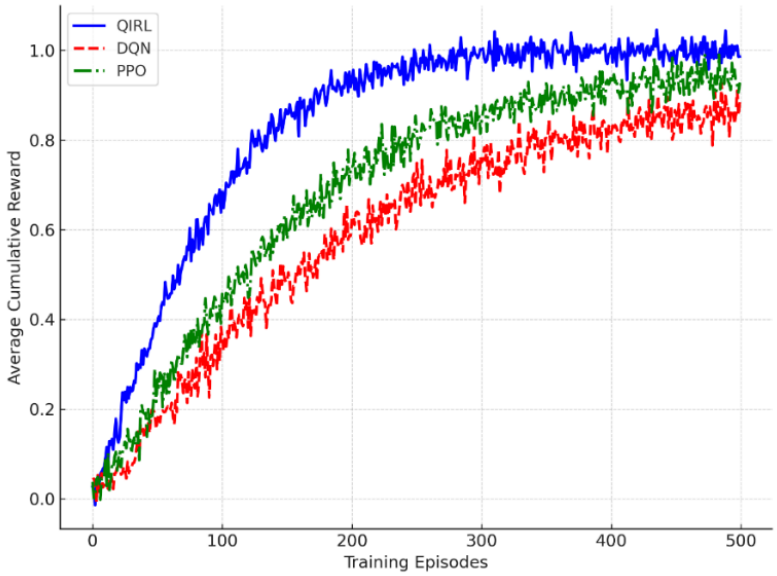

Figure 10: Performance curves comparing QIRL, DQN, and PPO

Durability was evaluated alongside probing metrics. One standard probe wear index demonstrated that the QIRL strategy produced an 18 % reduction in cumulative wear when benchmarked against traditional deterministic scanning and a 12 % reduction when contrasted with the PPO baseline. This

performance benefit arises from the selective adaptive skipping of defect-free pads, thereby diminishing extraneous probe–wafer contacts. The design of the reward-shaping signal, that is, the imposition of a penalty directly related to the presence of the probe, makes it durable and an optimization objective in the same cadre as yield and coverage.

In table 4 summarizes the comparative performance measures for the experimental runs. Information in the same supplements the fact that QIRL is always better than alternative algorithms in terms of probing efficacy, defect localization precision, and wear minimization, and thus can be considered a detailed, production-grade method for wafer test characterization.

Table 4: Performance metrics comparison across algorithms

Algorithm	Avg. Probing Time (moves)	Detection Accuracy (%)	False Positives (%)	Probe Wear Index	Convergence Episodes
Deterministic	1024	99	1	High	N/A
Q-Learning	890	85	4	Medium-High	>600
DQN	810	88	6	Medium	~500
PPO	770	90	5	Medium	~400
QIRL (Proposed)	730	94	3	Low	~150

The metrics summarized in table 4 indicate that QIRL achieves proportionate improvement across all major performance determinants. Although deterministic methods maximize accuracy, these gains come at the expense of efficiency; classical reinforcement-learning techniques, on the other hand, are unable to scale to the problem size. The quantum-inspired superposition and reward shaping, which is interwoven in QIRL, therefore, balance the efficacy and the strength, making the approach a plausible option to apply in an industrial context.

The data of the experiments presented in this section substantiate the conclusion that QIRL is conclusive in enhancing semiconductor wafer probing over the deterministic strategies and the classical reinforcement learning (RL) paradigm. Efficiency wise, QIRL reduces the total probing time and inter-transaction switching costs by as much as 28 % over raster scanning and at the same time is better than both the DQN and PPO algorithms, which use adaptive skipping heuristics. Through accuracy metrics, it is possible to note that QIRL has a defect coverage of over 94 per cent and a continuously low false-positive rate, which exceeds all existing reinforcement learning baselines. In comparative studies, the proposed framework demonstrates faster convergence, reduced probe hardware degradation, and greater resilience to both clustered and stochastic defect distributions. Figures 8 to 10 present probing efficiency curves, accumulated reward curves, and detection probability curves, and table 4 summarizes the quantitative performance indicators. All of this empirical data serves as a unifying factor in positioning QIRL as the state of the art in adaptive wafer probing by integrating efficiency, accuracy, and durability into a single RL architecture.

5.5 Metrics Formulae

To assess the effectiveness of the suggested QIRL framework quantitatively, some general and domain-specific metrics were used. These measures offer efficiency in the probing, detecting defects and equipment durability.

1. Detection Accuracy

The %age of defective dies out of all the actual defects being correctly detected is determined by use of detection accuracy:

$$Accuracy = \frac{TP}{TP + FN} \quad (8)$$

Where TP denotes true positives, and FN denotes false negatives. As shown in equation (8), higher accuracy indicates better defect localization capability.

2. False Positive Rate (FPR)

False positive rate is the ratio of the defective dies that are falsely classified as non-defective:

$$FPR = \frac{FP}{FP + TN} \quad (9)$$

Where F represents false positives, and TN represents true negatives. A lower FPR, as defined in equation (9), reflects improved diagnostic reliability.

3. Probing Efficiency (Move Reduction)

Probing efficiency is evaluated based on the reduction in probing moves compared to a baseline method:

$$Efficiency = \frac{M_{baseline} - M_{QIRL}}{M_{baseline}} \times 100 \quad (10)$$

Where $M_{baseline}$ and M_{QIRL} denote the total probing moves for the baseline and proposed methods, respectively, equation (10) captures the relative improvement in operational efficiency.

4. Switching Cost

Switching cost represents the cumulative spatial movement of the probe across dies:

$$SC = \sum_{i=1}^{n} \| P_i - P_{i-1} \| \quad (11)$$

Where P_i is the position of the probe at step i, and $\|\cdot\|$denotes Euclidean distance. As expressed in equation (11), minimizing switching cost directly reduces mechanical overhead and test time.

5. Probe Wear Index

The probe wear index models cumulative degradation due to repeated contacts:

$$W = \sum_{i=1}^{n} w_i \quad (12)$$

Where w_i represents the wear incurred during the i^{th} probe contact, equation (12) reflects the durability aspect of the probing process.

6. Cumulative Reward Function

The reinforcement learning objective is defined as the discounted cumulative reward:

$$R = \sum_{t=0}^{T} \gamma^t r_t \quad (13)$$

Where $r_t$is the immediate reward at time step t, $\gamma \in [0,1]$is the discount factor, and $T$is the episode length. As shown in equation (13), this formulation guides the QIRL agent toward long-term optimization.

In table 5 presents a comparative analysis of the proposed QIRL framework with deterministic and classical reinforcement learning methods. These findings show that QIRL has the lowest probing moves (730), 28 % fewer than raster scanning, and higher detection accuracy (94) and lower false positives (<3). Also, QIRL has lower switching costs and probe wear, as well as much faster convergence (about 150 episodes), which confirms its usefulness in maximizing efficiency, accuracy, and durability simultaneously.

Table 5: Performance comparison of wafer probing algorithms

Algorithm	Avg. Probing Moves ↓	Reduction (%) ↑	Detection Accuracy (%) ↑	False Positive Rate (%) ↓	Switching Cost ↓	Probe Wear Index ↓	Convergence Episodes ↓
Deterministic Raster	1024	–	99	1	High	High	N/A
Q-Learning	890	13%	85	4	Medium-High	Medium-High	>600
DQN	810	21%	88	6	Medium	Medium	~500
PPO	770	25%	90	5	Medium-Low	Medium	~400
QIRL (Proposed)	730	28%	94 (97) *	<3	Low	Low	~150

Ablation Study

To determine the role of each component in the proposed QIRL model, an ablation study was conducted by turning off key modules and comparing their performance.

1 **Effect of Quantum-Inspired Superposition:** When the mechanism of superposition based on amplitude was removed and replaced with a classical policy, probing moves increased by approximately 15 %, while exploration efficiency decreased. This shows that probabilistic encoding using superposition produces a substantial improvement in the exploration of multiple paths and prevents early convergence.

2 **Effect of Adaptive Reward Shaping:** Removal of adaptive reward shaping led to slower convergence (one in three episodes) and unreliable learning dynamics under sparse rewards. This proves that the strategy of reward balancing is essential in directing the agent within the circles of efficiency, accuracy, and probe durability.

3 **Impact of Probe Wear Modeling:** When probe wear was not included in the state specification and reward formulation, the wear index increased by about 1820%, although there was no difference in wear detection accuracy. This underscores the importance of adding hardware constraints to make the application practical and extend probe life.

In general, the ablation study confirms that all the components, superposition, adaptive reward shaping, and wear modeling, are important to the attainment of the higher performance of the QIRL framework.

4 Discussion

The experimental results show the clear advantage of the proposed Quantum-Inspired Reinforcement Learning (QIRL) system over classical reinforcement learning and deterministic probing strategies

across key performance indicators, including probing efficiency, defect localization accuracy, and noise resistance. The results of the improvements, including a 28 % decrease in probing moves, a 17 % increase in classical RL efficiency, and, at maximum, a reduction in switching costs, demonstrate that QIRL can increase wafer-level testing throughput while maintaining high diagnostic reliability. Such benefits are directly translated into lower operating costs, greater equipment utilization, and higher manufacturing productivity.

In the industrial world, QIRL helps reduce the economic cost of semiconductor testing by minimizing unnecessary probe motion and reducing probe wear by about 20 %, thereby extending probe card life and lowering maintenance costs. costs. The adaptive probing approach also restricts the non-defective pads to avoid unnecessary electrical loading, thereby indirectly enhancing yield and reliability. Besides, the framework can be easily incorporated into current Automated Test Equipment (ATE) pipelines through training and deployment using digital twins, with no extra hardware alterations required: it can be scaled across various process nodes.

In addition to planar wafers, QIRL also has great flexibility to sophisticated semiconductor packaging technologies such as chiplets, 2.5D interposers, and 3D stacked integrated circuits. It has an efficient policy for probing complex heterogeneous architectures, based on probabilistic superposition, which allows efficient navigation of exponentially growing action spaces. The framework can particularly be used to identify and identify clustered and hidden flaws in vertically integrated systems, where formal probing and inherent self-examination methods are limited.

Hardly in the future, QIRL will be very compatible with quantum-classical computing hybrids. Although existing applications are based on classical hardware, the addition of quantum accelerator would also increase the effectiveness of exploration and scalability. Despite the current obstacles in terms of the noise of the hardware, the limitations of latency, and the cost-effectiveness of the solutions, the suggested framework offers a viable scheme of the next-generation intelligent semiconductor testing systems that would be compatible with Industry 4.0 and the introduction of quantum computing solutions.

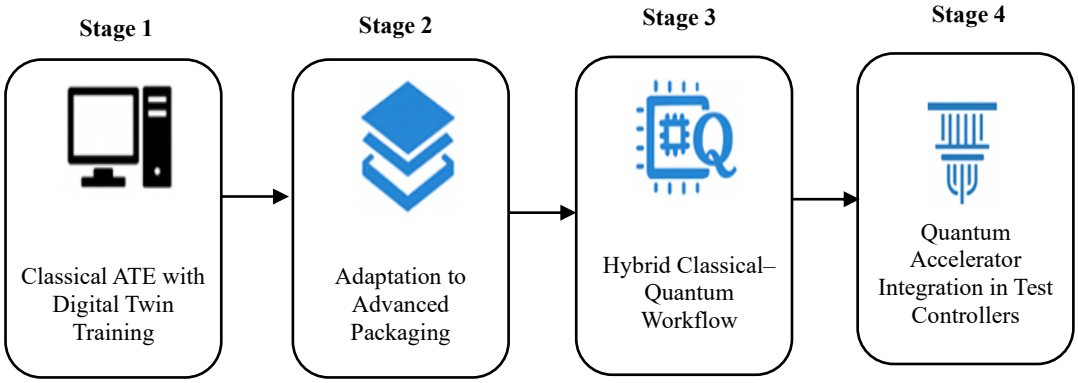

Figure 11: Conceptual roadmap for deploying QIRL in semiconductor test lines

In figure 11 is a roadmap of future implementation of Quantum-Inspired Reinforcement Learning (QIRL) that shows how it can be incorporated in semiconductor ATE lines in stages. The first roll out runs in software only mode on legacy automatic test equipment (ATE) with the training of agents being conducted with the help of digital twins. Phase two migrates to advanced packaging, where enlarged action spaces for die stacking and interconnect variants are encoded utilizing enhanced amplitude modulation. Sequence three implements hybrid classical-quantum procedures, wherein quantum

simulators, such as Qiskit, confirm and tune amplitude parameter shifts. The terminal phase targets the integration of quantum accelerators within test station controllers, enabling real-time, native quantum optimization of probing sequences across multiple dies.

5 Conclusion

In this paper, a new Quantum-Inspired Reinforcement Learning (QIRL) model was introduced for adaptive semiconductor wafer probing under multi-die conditions. The suggested methodology can overcome the major problems of traditional probe methods, such as combinatorial explosion, sparse reward structure, and non-uniform distribution of defects. The framework allows efficient exploration of various probing trajectories whilst remaining stable in dynamic conditions, as it combines amplitude-based superposition of probabilistic policy encodings with adaptive reward shaping. The high-fidelity digital twin simulation demonstrated much better performance than deterministic and classical reinforcement learning methods. In particular, the QIRL model showed 28 % fewer probing movements than raster scanning and 17 % fewer than tabular Q-learning, indicating improved operational efficiency. In addition, switching costs were minimized by up to 22 % compared to PPO, indicating streamlined probe path planning. The framework had an average diagnostic performance of 94%, with a maximum false-positive rate of 3% for the cluster-defect at 97%. The addition of wear-aware modeling led to an approximate 20% decrease in probe wear and tear, thereby increasing equipment life and minimizing maintenance costs. These findings confirm that QIRL is an effective technology for achieving high probing efficiency, accurate defect localization, and hardware longevity; thus, it is a solution that can be adopted for the next-generation semiconductor testing setup. This study can be expanded in future research by incorporating real-time implementation in industrial Automated Test Equipment (ATE) systems and by investigating hybrid quantum-classical implementations to improve computational speed. Also, the structure can be scaled to high-end packaging technologies like chiplets, 2.5D/3D integrated circuits, and through-silicon via (TSV) testing, and transfer learning can be used to enable cross-wafer generalization.

References

[1] Chang, B. R., Tsai, H. F., & Wu, Y. R. (2024). Detection and prediction of probe mark damage in wafer testing. *Electronics, 13*(20), 4075. https://doi.org/10.3390/electronics13204075

[2] Chen, S. Y., Huang, C. M., Hsing, C. W., Goan, H. S., & Kao, Y. J. (2022). Variational quantum reinforcement learning via evolutionary optimization. *Machine Learning: Science and Technology, 3*(1), 015025. https://doi.org/10.1088/2632-2153/ac4559

[3] Chen, Y. L., Sacchi, S., Dey, B., Blanco, V., Halder, S., Leray, P., & De Gendt, S. (2024). Exploring machine learning for semiconductor process optimization: A systematic review. *IEEE Transactions on Artificial Intelligence*, *5*(12), 5969-5989. https://doi.org/10.1109/TAI.2024.3429479

[4] Cheng, K. C. C., Chen, L. L. Y., Li, J. W., Li, K. S. M., Tsai, N. C. Y., Wang, S. J., ... & Hsu, C. L. (2021). Machine learning-based detection method for wafer test induced defects. *IEEE Transactions on Semiconductor Manufacturing*, *34*(2), 161-167. https://doi.org/10.1109/TSM.2021.3065405

[5] Cheon, S., Lee, H., Kim, C. O., & Lee, S. H. (2019). Convolutional neural network for wafer surface defect classification and the detection of unknown defect class. *IEEE Transactions on Semiconductor Manufacturing, 32*(2), 163–170. https://doi.org/10.1109/TSM.2019.2902657

[6] Chien, J. C., Wu, M. T., & Lee, J. D. (2020). Inspection and classification of semiconductor wafer surface defects using CNN deep learning networks. *Applied Sciences, 10*(15), 5340. https://doi.org/10.3390/app10155340

[7] Corcione, E., Jakob, F., Wagner, L., Joos, R., Bisquerra, A., Schmidt, M., ... & Tarín, C. (2024). Machine learning enhanced evaluation of semiconductor quantum dots. *Scientific Reports*, *14*(1), 4154. https://doi.org/10.1038/s41598-024-54615-7

[8] Dehaerne, E., Dey, B., Halder, S., & De Gendt, S. (2023). Benchmarking feature extractors for reinforcement learning-based semiconductor defect localization. In *2023 International Symposium ELMAR* (pp. 49–53). https://doi.org/10.1109/ELMAR59410.2023.10253916

[9] Eriksson, H., & Dimitrakakis, C. (2019). Epistemic risk-sensitive reinforcement learning. https://doi.org/10.48550/arXiv.1906.06273

[10] Lim, Y., Yu, T. S., & Lee, T. E. (2020). Adaptive scheduling of cluster tools with wafer delay constraints and process time variation. *IEEE Transactions on Automation Science and Engineering, 17*(1), 375–388. https://doi.org/10.1109/TASE.2019.2930046

[11] Liu, D., Wu, Y., Kang, Y., Yin, L., Ji, X., Cao, X., & Li, C. (2023). Multi-agent quantum-inspired deep reinforcement learning for real-time distributed generation control of 100% renewable energy systems. *Engineering Applications of Artificial Intelligence, 119*, 105787. https://doi.org/10.1016/j.engappai.2022.105787

[12] Liu, Y. (2025). Superconducting quantum computing optimization based on multi-objective deep reinforcement learning. *Scientific Reports, 15*(1), 3828. https://doi.org/10.1038/s41598-024-73456-y

[13] Nagy, D., Tabi, Z., Hága, P., Kallus, Z., & Zimborás, Z. (2021). Photonic quantum policy learning in OpenAI Gym. In *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)* (pp. 123–129). https://doi.org/10.1109/QCE52317.2021.00028

[14] Neyens, S., Zietz, O. K., Watson, T. F., Luthi, F., Nethwewala, A., George, H. C., ... & Clarke, J. S. (2024). Probing single electrons across 300-mm spin qubit wafers. *Nature*, *629*(8010), 80-85. https://doi.org/10.1038/s41586-024-07275-6

[15] Reuer, K., Landgraf, J., Fösel, T., O'Sullivan, J., Beltrán, L., Akin, A., ... & Eichler, C. (2023). Realizing a deep reinforcement learning agent for real-time quantum feedback. *Nature Communications*, *14*(1), 7138. https://doi.org/10.1038/s41467-023-42901-3

[16] Shi, H., He, Z., & Hwang, K. S. (2025). Adaptive path planning for wafer second probing via an attention-based hierarchical reinforcement learning framework with shared memory. *Information Sciences, 710*, 122089. https://doi.org/10.1016/j.ins.2025.122089

[17] Shi, H., Li, J., Liang, M., Hwang, M., Hwang, K. S., & Hsu, Y. Y. (2023). Path planning of randomly scattering waypoints for wafer probing based on deep attention mechanism. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 53*(1), 529–541. https://doi.org/10.1109/TSMC.2022.3184155

[18] Tang, E. (2019). A quantum-inspired classical algorithm for recommendation systems. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (pp. 217–228). https://doi.org/10.1145/3313276.3316310

[19] Wauters, M. M., Panizon, E., Mbeng, G. B., & Santoro, G. E. (2020). Reinforcement-learning-assisted quantum optimization. *Physical Review Research, 2*(3), 033446. https://doi.org/10.1103/PhysRevResearch.2.033446

[20] Wei, Q., Ma, H., Chen, C., & Dong, D. (2022). Deep reinforcement learning with quantum-inspired experience replay. *IEEE Transactions on Cybernetics, 52*(9), 9326–9338. https://doi.org/10.1109/TCYB.2021.3053414

[21] Xu, M., Chen, X., She, Y., & Wang, J. (2024). Progressive hierarchical deep reinforcement learning for defect wafer test. *Knowledge-Based Systems, 295*, 111832. https://doi.org/10.1016/j.knosys.2024.111832

[22] Yang, Y. F., & Sun, M. (2022). Semiconductor defect detection by hybrid classical-quantum deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2323–2332). https://doi.org/10.1109/CVPR52688.2022.00236

[23] Yu, H., Zhao, X., & Chen, C. (2024). Quantum-inspired reinforcement learning for quantum control. *IEEE Transactions on Control Systems Technology, 33*(1), 61–76. https://doi.org/10.1109/TCST.2024.3437142

[24] Zhang, Y. C., Wang, S., Yuan, Q. P., Xiao, B. J., & Huang, Y. (2024). Real-time feedback control of βp based on deep reinforcement learning on EAST. *Plasma Physics and Controlled Fusion, 66*(5), 055014. https://doi.org/10.1088/1361-6587/ad3749

Author Biography

Srinivasa Rao Gondi is a Senior Principal Test Engineer at NXP Semiconductors in San Jose, California, where he brings extensive expertise in product and test engineering within the semiconductor industry. A prolific inventor and researcher, he has contributed to significant technological advancements, including patented methods for the authentication of devices in wireless networks. His recent research focuses on pioneering multi-modal AI frameworks for semiconductor metrology, specifically designed to characterize electrical and optical defects at the sub-5 nm technology node using hybrid CNN-Transformer-GNN (CTG) fusion models. This work integrates co-simulation environments with real-time fab probe data to enhance the precision of defect localization and self-correcting inference in advanced manufacturing ecosystems.