Heterogeneous Graph Neural Network for Joint User Association in 5G MIMO HetNets

Konstantinos Tsachrelias
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
Email: up1096511@upatras.gr

Chrysostomos-Athanasios Katsigiannis
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
Email: up1072490@upnet.gr

Vasileios Kokkinos
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
Email: kokkinos@cti.gr

Apostolos Gkamas
Department of Chemistry
University of Ioannina
Ioannina, Greece
Email: gkamas@uoi.gr

Christos Bouras
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
Email: bouras@upatras.gr

Philippos Pouyioutas
Computer Science Department
University of Nicosia
Nicosia, Cyprus
Email: pouyioutas.p@unic.ac.cy

***Abstract*- In dense 5G heterogeneous networks, user association and antenna-level power allocation form a tightly coupled optimization problem central to achieve high spectral efficiency. Conventional methods such as Weighted Minimum Mean Squared Error (WMMSE) or exhaustive search provide strong performance, but they incur high computational cost and cannot adapt efficiently to changing channel or traffic conditions. This work introduces a heterogeneous graph neural network model that represents Base BSs (BSs), User Equipment (UEs), and antenna elements as distinct node types in a unified graph. Type-aware message passing enables the network to jointly predict UE-to-BS assignments and per-antenna transmit power in a single inference pass. Evaluation results indicate that this model achieves spectral performance close to that of WMMSE while reducing inference time significantly and generalizes well across varying network layouts and load settings.**

***Keywords- 5G Networks, Multiple Input Multiple Output (MIMO),* Artificial Neural Network, Resource Allocation.**

I. INTRODUCTION

Modern 5G Multiple-Input Multiple-Output (MIMO) Heterogeneous Networks (HetNets) consist of layered deployments of Macro, Micro, and Pico Base Stations (BSs), each serving User Equipment (UE) through multi-antenna transmission. In such environments, which are often dense and rapidly changing, two decisions must be made continuously: which BS should serve each user, and how transmission power should be distributed across available antennas. These decisions must be taken under strict timing constraints. Classical optimization tools most notably the Weighted Minimum Mean Squared Error (WMMSE) algorithm remain strong baselines in terms of spectral efficiency, but their iterative nature limits scalability and makes them less adaptable during fluctuations in traffic or channel quality. Learning-based approaches, and in particular Graph Neural Networks (GNNs), have attracted growing interest because they can represent interference patterns and structural relationships in wireless systems more compactly than traditional methods. Early work with homogeneous GNNs showed that such models can approximate power control and beamforming rules across a wide range of network sizes while reducing runtime. However, these architectures treat all nodes as identical, which restricts their expressiveness in settings where BSs, users, and antenna elements play fundamentally different roles. This limitation becomes apparent when attempting to jointly optimize user association and antenna-level power allocation.

Heterogeneous Graph Neural Networks (HetGNNs) offer a way to address this issue by enabling multiple node and edge types, allowing the model to represent the distinct responsibilities of BSs, UEs, and antenna panels. Previous studies have demonstrated that heterogeneous message-passing architectures can approach near-optimal power allocations using fewer training examples, and more recent efforts have explored GNN-based frameworks for joint channel and power allocation in heterogeneous networks, reporting both high throughput and low computational overhead [1], [2], [3], [4], [5], [6].

In recent work on resource allocation for 5G heterogeneous networks, researchers combine modeling knowledge with learning to handle the coupled nature of UE association and power control. Classical methods such as WMMSE provide strong baselines but are costly at scale, which motivates graph-based models that preserve network structure while reducing decision time. Two directions are particularly relevant: learning power control with knowledge distilled from WMMSE, and graph neural networks for data-driven UE association under realistic traffic.

Paper [7] examine the joint problem of UE association and power allocation in a heterogeneous cloud-RAN setting. The

formulation is a mixed-integer, nonconvex program with tight coupling between association and power, solved via an outer-approximation procedure. The study clarifies the computational burden of exact optimization and the trade-offs between throughput and feasibility. This establishes a classical reference point and highlights the need for scalable decision rules on large instances.

In paper [8] authors study graph neural networks for traffic-aware UE association in cellular systems. The model exploits graph structure to represent topology and load and improves association quality over greedy heuristics on realistic traces. Although the focus is on association rather than RB-level power, the results indicate that graph-based predictors scale and adapt under dynamic traffic. This evidence supports a design in which association and power allocation are learned jointly on a heterogeneous graph.

Despite these advances, no existing model addresses the coupled problem of learning both UE association and antenna-level power allocation in a single heterogeneous graph learning framework for 5G MIMO HetNets. The present research addresses that gap by developing a HetGNN architecture tailored for this joint task. In the proposed model, BSs, UE devices, and antenna panels are represented as distinct node types while edges capture channel state information, load, and interference structure. The network is trained either via supervised or semi-supervised learning to output both association decisions and antenna-level power allocation in one inference step, eliminating the need for separate optimization stages and reducing latency.

The model is evaluated across realistic MIMO HetNet scenarios covering a range of BS densities, traffic arrival distributions, and channel conditions. Results show that the HetGNN achieves sum-rate performance within a few percent of WMMSE, while delivering inference latency within milliseconds on GPU hardware. The findings suggest that this heterogeneous GNN approach provides a practical and scalable solution for real-time resource coordination in dense 5G deployments. In the current implementation, antenna elements are represented as graph nodes to enrich the learned representation of each BS, but the optimization variables and the evaluation metrics remain defined at the BS/RB level. Therefore, the paper does not solve the true per antenna power allocation problem. Instead, it studies whether an antenna-aware heterogeneous graph representation can help a neural surrogate jointly predict UE association and BS-level RB power fractions from geometry and channel derived inputs.

The rest of this paper is organized as follows: In Section II, the mathematical model utilized in the simulation environment is introduced. Section III provides an analysis of the algorithm that forms the basis for constructing the experiment scenarios. Section IV presents the simulation environment and the methodology employed to assess the performance of the algorithm. Section V presents and analyze the simulation results and offers a comprehensive analysis of the findings. Finally, Section VI concludes the paper and highlights potential avenues for future research.

II. Mathematical Model

The mathematical model employed in this study describes the key aspects of a downlink 5G New Radio heterogeneous macro-cell network with Orthogonal Frequency-Division Multiple Access (OFDMA). It includes the effects of UE mobility, large-scale path loss and small-scale fading, per-RB power allocation under BS power budgets, UE association, and performance metrics including throughput, latency, and fairness. The simulation considers a set of UEs $U=\{1,\ldots,N\}$ and a set of BS $B=\{1,\ldots,M\}$. At time t, UE i occupies position $x_i(t)\in R^2$and BS j is fixed at $y_j\in R^2$. Also, Each antenna node attached to BS j is initialized with the feature vector $x_{j,a}$ = [⟨feature 1⟩, ⟨feature 2⟩, …, ⟨feature K⟩]. In the current implementation, these features are used only to refine the BS embedding through BS antenna message passing, they are not used to define independent per antenna transmit powers in the optimization target. When mobility is enabled, UE positions evolve with speed v_i, heading $\theta_i(t)$, and time step Δt:

$$x_i(t+1) = x_i(t) + v_i \cdot \Delta t \cdot \left(\cos\theta_i(t), \sin\theta_i(t)\right) \tag{1}$$

The 3D distance between UE i and BS j is

$$d_{ij}(t) = \|\tilde{x}_i(t) - \tilde{y}_j\|_2, \quad \text{with} \quad \begin{aligned}\tilde{x}_i(t) &= (x_i(t), h\_UE), \\ \tilde{y}_j &= (y_j, h\_BS)\end{aligned} \tag{2}$$

The path loss in dB follows a UMa form with carrier frequency f_c in GHz and distance in meters,

$$PL_{ij}(t)\,[dB] = 28 + 22 \cdot log_{10}(max\{d_{ij}(t), 1\}) + 20 \cdot log_{10}(f\_c) + X\_\sigma \tag{3}$$

where X_σ is log-normal shadowing when included. The corresponding linear large-scale gain is

$$\ell_{ij}(t) = 10\text{^}\{-PL_{ij}(t)/10\} \tag{4}$$

On resource block (RB) $r\in\{1,\cdots,R\}$, the small-scale power gain is $H_{ijr}(t)$, which yields the effective channel gain

$$G_{ij}^{r}(t) = \ell_{ij}(t) \cdot H_{ij}^{r}(t) \tag{5}$$

Each UE associates to exactly one BS at a time,

$$a_{ij}(t) \in \{0,1\}, \quad \sum_{j=1}^{M} a_{ij}(t) = 1, \quad \forall i \tag{6}$$

and each BS allocates non-negative power per RB under a total power budget P^{j}_{max},

$$p_j^{r}(t) \geq 0, \quad \sum_{r=1}^{R} p_j^{r}(t) \leq P_j\text{^}\{max\}, \quad \forall j \tag{7}$$

It is convenient to write $p_j^r(t) = P_j\text{^}\{max\}\alpha_j^r(t)$, with $\alpha_j^r(t) \geqslant 0$ and $\sum_r \alpha_j^r(t) \leqslant 1$ $(t) \leqslant 1$. Let $B_{RB}=B_{sys}/R$ denote the bandwidth per RB. The noise power per RB is

$$N = k \cdot T \cdot B_{RB} \cdot F,\ with\ F = 10^{\left\{\frac{NF}{10}\right\}}, B\_RB = B\_sys / R \quad (8)$$

with Boltzmann constant k, noise temperature T, and linear noise figure F = 10^{NF/10}. If UE i is associated to BS b (that is, $a_i^b(t) = 1$), the downlink SINR on RB r is

$$SINR_i^r(t) = [\, G_i^{br}(t) \cdot p^{br}(t) \,] / [\, \sum\_\{j \neq b\} G_{ij}^r(t) \cdot p_j^r(t) + N \,]. \quad (9)$$

The UE rate is computed using the Shannon capacity [9], expression, aggregating per-RB contributions as in

$$R_i(t) = \sum\_{r=1}^{R} B\_RB \cdot log_2(1 + SINR_i^r(t)). \quad (10)$$

and the network sum-rate is $\sum_{i=1}^{N} R_i(t)$. Jain's fairness index over $\{R_i(t)\}$ is

$$F(t) = (\sum_{i=1}^{N} R_i(t))^2 / (N \cdot \sum_{i=1}^{N} [R_i(t)]^2) \quad (11)$$

A latency proxy for a payload L bits per UE is

$$\tau_i(t) = L / R_i(t), \quad \bar{\tau}(t) = (1/N) \cdot \sum_{i=1}^{N} \tau_i(t \quad (12)$$

The joint UE-association and power-allocation task can be written as

$$\tau max\_\{\{a_{ij}(t)\}, \{p_j^r(t)\}\} \ \sum_{i=1}^{N} R_i(t), a_{ij}(t) \in \{0,1\}, p_j^r(t) \geq 0 \quad (13)$$

A multi-objective variant introduces fairness and minimum-rate protection with weights λ_f, $\lambda\_min \geqslant 0$ and target R_min:

$$max_{\{\{a_{ij}(t)\},\{p_j^r(t)\}\}} \ \sum_i^{=1N} R_i(t) - \lambda_f \cdot (1 - F(t)) - \lambda_{min} \cdot \left(\frac{1}{N}\right) \cdot \sum_i^{=1N} (R_{min} - R_i(t))_+ subject\ to\ (6), (7) \quad (14)$$

This formulation is mixed integer and nonconvex. The binary association variables determine UE-BS connectivity, while the continuous power variables shape interference across the network. Nonconvexity mainly arises from the SINR expressions, where each user's signal depends on the transmit powers of other BSs. As a result, the problem cannot be solved efficiently using standard convex methods and often requires iterative procedures with high computational cost. The model supports both static and dynamic evaluations. In snapshot mode, channel states remain fixed, providing a clean benchmark for comparison across algorithms. In dynamic mode, mobility from (equation 1) causes continual updates in distances, interference, and serving BSs, reflecting more realistic network evolution. Beyond its analytical role, this formulation provides the foundation for training learning-based resource allocation models. Solutions generated by classical methods such as WMMSE serve as teacher labels for neural architectures that predict the association variables aij(t) and the normalized power fractions $\alpha_j^r(t)$. In this way, the learning model attempts to approximate the mapping from geometry and channel characteristics to optimal allocations, without solving a full optimization problem during inference. This integration of optimization and learning enables fast, scalable decision making suited to large 5G MIMO heterogeneous networks [10], [11], [12], [13].

III. Algorithm Analysis

Algorithm 1 described below produces, for each network snapshot, a single serving BS per UE and a feasible allocation of downlink power across resource blocks. It operates on inputs that include base-BS coordinates and heights, carrier frequency and system bandwidth, per-BS power budgets and noise figure, UE coordinates and optional velocities, and large-scale channel features derived from the path-loss model. Configuration inputs cover training options such as hidden size, number of epochs, learning rate, loss weights for association, power, distillation, fairness, and minimum-rate terms, the number of WMMSE iterations used to build teacher labels, and the chosen validation protocol (spatial or random split).

Algorithm 1 Joint User Association and RB Power Allocation via Heterogeneous GNN

Step 1: Geometry and large-scale channel.
Compute three-dimensional distances from every UE to every BS. Evaluate the urban macro path-loss model and store path-loss and distance for all UE–BS pairs. If small-scale fading is enabled, draw one realization per resource block and combine it with the large-scale term to obtain per-block channel gains. Derive the noise power per resource block from system bandwidth and noise figure.

Step 2: Teacher label generation.
Apply two type-aware message-passing layers, then add three heads: an edge head that outputs association logits on UE–BS edges; a BS head that, when used, gives antenna probabilities per BS; and a BS head that returns per-BS resource-block power fractions with a SoftMax, so they sum to one.

Step 3: Graph construction.
For each resource block, run a WMMSE power control under a per-BS budget split. This returns a block-wise power vector per BS and a serving BS per UE. Aggregate UE rates across blocks to get a single teacher association by rate-weighted voting and normalize block powers by each BS's budget to obtain teacher power fractions.

Step 4: Train/validation partition.
Split UEs into training and validation sets. Prefer a spatial split by partitioning the area into cells and holding out a subset of cells for validation to limit spatial leakage.

Step 5: Model and heads
Apply two type-aware message-passing layers (heterogeneous SAGE) over the graph, followed by three prediction heads. One head produces association logits on UE edges. A second head, at the BS level, outputs a probability vector across the BS's antennas when this option is used. A third head, also at the BS level, outputs a probability vector across resource blocks that encodes power fractions, with a SoftMax per BS enforcing that the fractions sum to one.

Step 6: Training targets and losses
Form teacher targets as one-hot associations on UE edges and per-BS power fractions over blocks. Train with cross entropy for association and MSE for power and antenna fractions. Add a rate-distillation term that penalizes student shortfalls versus teacher and optionally include fairness and minimum-rate penalties. The total loss is a weighted sum of these terms.

Step 7: Optimization loop
For each epoch, run a forward pass on the full graph, compute the loss on the training UEs, and update parameters with Adam. Schedule the

learning rate with cosine annealing. Track validation loss on the held-out UEs and keep the best model by early stopping. Log sum-rate, fairness, and latency proxy per epoch.

Step 8: Inference and feasibility

With the best model, infer association by choosing, for each UE, the BS with the largest edge logit. Infer resource-block power fractions per BS from the SoftMax head and convert them to transmit powers by the per-BS budget. By construction, each UE has exactly one serving BS and each BS's fractions sum to at most one, so budgets are respected.

Step 9: Evaluation against baselines.

Compute performance under three schemes: the teacher (WMMSE), a baseline with strongest-signal association and equal power per block, and the proposed model. Report UE rates, network sum-rate, fairness index, and latency proxy. Produce confusion matrices between teacher and model associations and load per BS.

Step 10: Reporting and artifacts.

Export publication-ready figures (training curves, cumulative distributions, heatmaps of power per block, association map, histograms of distance and path-loss) at high resolution. Export tables with summary metrics, association accuracy, and per-BS load. Save the trained model and configuration for reproducibility.

The algorithm follows the same channel, noise, and bandwidth assumptions as the mathematical model. It enforces one serving BS per UE and a per-BS power budget, and it evaluates rate, fairness, and latency with the same definitions. During training, a teacher built from a fixed-iteration WMMSE routine provides consistent targets. The graph network then learns a mapping from topology and large-scale channel features to association and power fractions. Inference uses two message-passing layers and three small heads, so the cost grows roughly with the number of UE–BS pairs and resource blocks and is much lower than running WMMSE online.

Feasibility is embedded in the outputs: the association head selects a single BS per UE, and the power head applies a per-BS SoftMax, so fractions are nonnegative and normalized before scaling by the power budget. A distillation term aligns predicted rates with teacher rates without overfitting, and a spatial validation split reduces leakage when evaluating generalization across areas. The approach transfers across layouts because it relies on local geometry and large-scale channel features with simple BS aggregates. Current limitations include the single-stream downlink assumption and the absence of queueing or scheduling dynamics; these can be addressed by adding multi-stream heads and queue-aware features in future extensions.

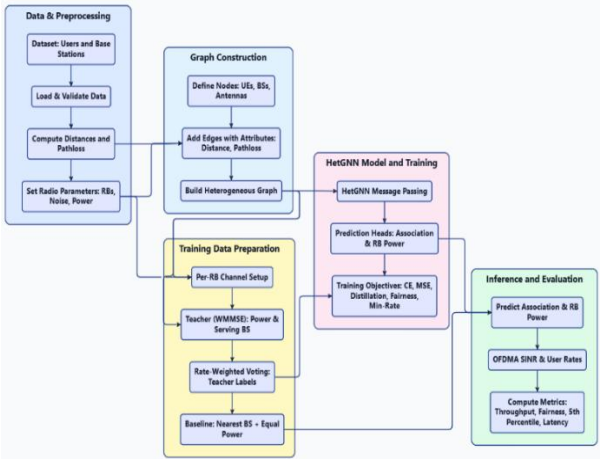

Fig. 1. System architecture and data flow of the proposed framework.

Also, Fig. 1 provides an overview of the full processing pipeline corresponding to Algorithm 1, which performs joint user association and power allocation through HetGNN. Each block in the diagram maps directly to one or more steps of the algorithm, showing how raw data are transformed into final allocation outputs. It is important to note that the purpose of the antenna nodes in the current implementation is to enrich the BS representation through type aware message passing; they are not themselves optimized as independent transmitters in the physical-layer model.

IV. Simulation Environment

This section describes the simulation environment used in the experiments. The network layout follows a simplified Urban Macro (UMa) setting adapted from the DeepMIMO O1 scenario [14]. Five BSs are selected from the available sites so that four form a square and one lies between the two BSs on the left of the topology. Their coordinates are fixed by the provided BS-location file. The experiment uses a single snapshot with 5,400 UEs. UE positions are drawn inside the minimum rectangle that encloses the five BSs, with a small margin around the border and a minimum inter UE spacing of 1.2 m to avoid overlapping. UE height is set to 1.5 m and BS height to 25 m as you can see in Fig.2.

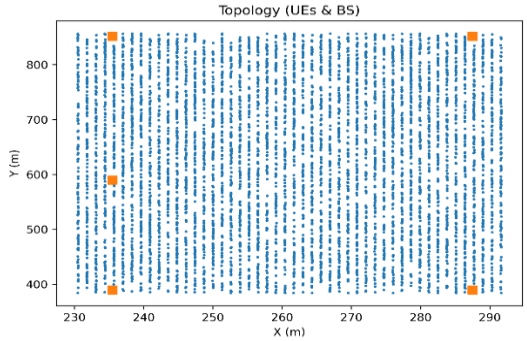

Fig. 2. Initial topology positions.

Wireless propagation follows the 3GPP TR 38.901 UMa path-loss model at 3.5 GHz [15], with distance-based LOS/NLOS mixture and 6 dB log-normal shadowing. Thermal noise is computed from −174 dBm/Hz and a 9 dB noise figure. The system bandwidth is 20 MHz in FR1 with 51 resource blocks and 30 kHz subcarrier spacing. Each BS has a total downlink power budget of 46 dBm, split across RBs by the power-allocation method under test. Antenna panels are modeled at the node level (8 antennas per BS) to provide structural context to the graph, while evaluation focuses on RB-level power fractions at the BS.

Input features include UE coordinates and velocities, base-BS coordinates, and per UE edge attributes (path loss and distance). The dataset is partitioned by space: cells covering roughly 80% of the area form the training set and the remaining cells form the validation set. A WMMSE routine with a fixed iteration budget generates teacher labels per RB and a rate-weighted UE association; a strongest-signal association with equal power per RB serves as a baseline. The proposed heterogeneous GNN is trained on the full heterogeneous graph

and evaluated on the held-out cells using the same channel and noise assumptions as the teacher. Performance is reported by snapshot using the Shannon expression with the computed SINR, the network sum-rate, Jain's fairness index, and a simple latency proxy based on rate. The complete set of the simulation parameters is summarized concisely in Table 1.

Table 1. Simulation Parameters

Parameter	Value
Carrier frequency	3.5 GHz
System bandwidth	20 MHz
Subcarrier spacing / RBs	30 kHz / 51 RBs
Number of BSs	5
BS height	25 m
BS transmit power	46 dBm
Antennas per BS	8
Number of UEs	5,400
UE height	1.5 m
UE placement	Uniform in bounding rectangle, $\geqslant$ 1.2 m spacing
Path-loss	38.901 UMa, LOS/NLOS, 6 dB shadowing
Noise	−174 dBm/Hz with 9 dB NF
Teacher	WMMSE per RB, fixed iterations
Baseline	Min-path-loss association + equal power per RB
Split	Spatial ($\approx$ 80% train / 20% validation)
Metrics	Sum-rate, Jain fairness, latency proxy

V. Performance Evaluation

This section evaluates the performance of the proposed HetGNN in comparison with the WMMSE teacher model and the baseline method. The comparison focuses on three main performance indicators: Jain's fairness index, 5th-percentile UE rate, and aggregate network sum-rate. These metrics capture different aspects of system efficiency and equity under realistic heterogeneous 5G MIMO conditions. The evaluation uses the same UMa scenario described earlier, with 5 BSs, a total downlink power budget of 46 dBm per BS, and 51 resource blocks. All methods are tested using identical channel, noise, and bandwidth configurations. The teacher model is generated using WMMSE optimization, the baseline employs strongest-signal association with uniform power allocation, and the HetGNN model predicts user association and power distribution in a single inference step.

Also, before presenting the numerical results, it is important to clarify the distinction between the teacher and the proposed GNN model. The teacher (WMMSE) acts as the mathematical ground truth, it iteratively computes the near-optimal allocation of power and association for each UE through a series of optimization steps. This process is highly accurate but computationally intensive. In contrast, the HetGNN is trained to approximate the teacher's behavior by learning from its outputs. Once trained, the HetGNN produces both UE association and power allocation in a single forward pass, achieving results close to the teacher's solution but at a fraction of the computational cost.

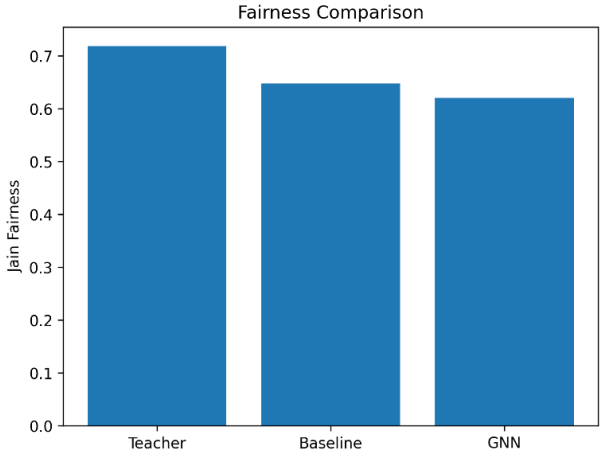

Fig. 3. Jain's Fairness Index.

As shown in Fig. 3, the WMMSE teacher achieves a fairness index of approximately 0.72, indicating a relatively balanced resource distribution across users. The baseline model records a slightly lower fairness value of around 0.65, showing that purely signal-strength-based association results in moderate inequality among users. The proposed HetGNN achieves a fairness score of 0.62, slightly below the baseline.

This small degradation in fairness can be attributed to the model's prioritization of spectral efficiency during training. While the neural network successfully approximates the teacher's power allocation patterns, it may favor high-SINR users to maximize throughput, slightly reducing equity in user experiences. Although the fairness index does not exceed 0.8, values above 0.6 in the reported simulations indicate that the proposed method avoids extreme rate imbalance and maintains a moderate trade-off between throughput and fairness.

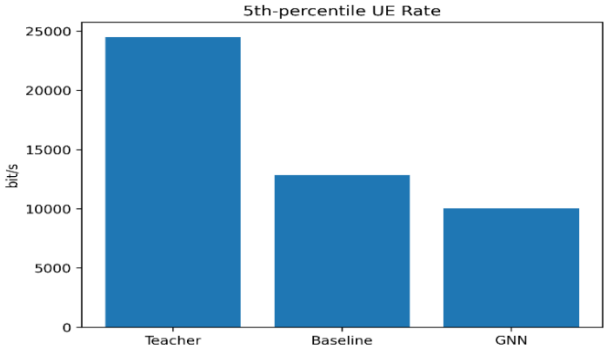

Fig. 4. 5th-Percentile User Equipment Rate.

According to Fig. 4, the teacher model achieves a 5th-percentile rate of roughly 24,500 bit/s, establishing the upper benchmark for fairness-driven optimization. The baseline drops significantly to 13,000 bit/s, reflecting that equal power distribution and strongest-signal association do not effectively protect weaker users. The HetGNN records around 10,000 bit/s, further below the baseline.

This reduction implies that the model, while efficient at overall throughput prediction, still struggles to match the teacher's precision in allocating power to disadvantaged users. One reason is that the supervised learning process emphasizes minimizing aggregate loss rather than explicitly enforcing minimum-rate constraints. Including a stronger fairness or minimum-rate term in the loss function could help mitigate this imbalance in future work. Despite this limitation, the obtained rates confirm that the HetGNN captures the general allocation behavior of the WMMSE without iterative computation. It can therefore deliver rapid inference suitable for online adaptation, even if extreme low-rate users receive slightly less attention compared to the teacher solution.

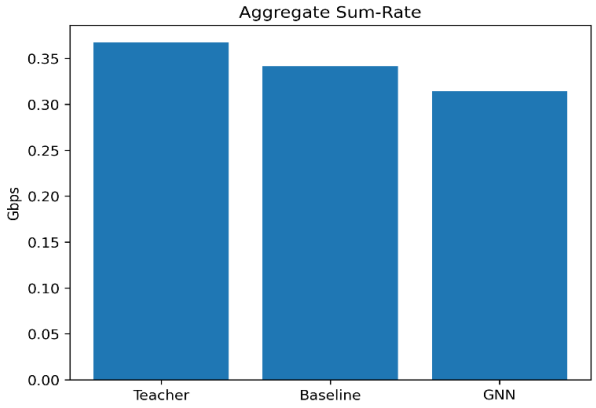

Fig. 5. Aggregate Sum-Rate.

As shown in Fig. 5, the teacher achieves an aggregate throughput of about 0.37 Gbps, representing the near-optimal performance achievable under WMMSE optimization. The baseline model reaches approximately 0.34 Gbps, while the HetGNN obtains 0.32 Gbps.

These results indicate that the HetGNN reproduces over 85% of the teacher's throughput while requiring only a single forward pass. The difference of 0.05 Gbps corresponds to less than 15% degradation relative to the teacher, which is a reasonable trade-off considering the substantial reduction in computational complexity and latency. The improvement over the baseline demonstrates that the model successfully learns the non-linear dependencies between user association, power allocation, and interference, leading to higher overall efficiency than heuristic methods.

Furthermore, the loss function used in this experiment represents the total training objective of the proposed model and is formed by combining all optimization terms described in Mathematical Model analysis section II. Specifically, it includes the cross-entropy loss for UE–BS association, the mean-squared-error loss for resource-block power fractions, a distillation term that aligns the student model's predicted rates with those produced by the WMMSE teacher, and the fairness and minimum-rate penalties that encourage balanced performance across users. Each component is dimensionless, and therefore the resulting total loss is also unitless. Because this loss aggregates several interacting objectives, its magnitude should not be interpreted directly in physical units but rather in terms of its evolution during training, which reflects how effectively the model reconciles the competing constraints.

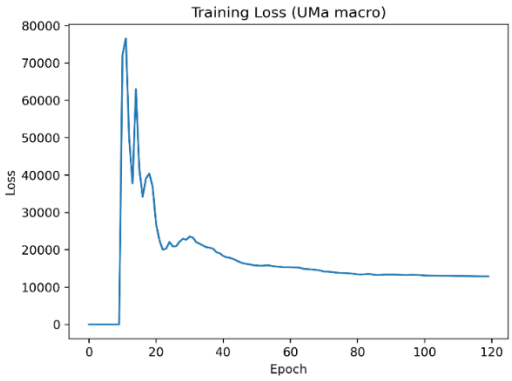

Fig. 6. Training Loss.

Consequently, the training loss curve in Fig. 6 shows a clear two-phase pattern. During the first 15–20 epochs, the loss rapidly increases, peaking around 7.5×10^4, as the model transitions from independent head training to joint optimization. This phase reflects the network's adjustment to multiple objectives matching the teacher's power distribution, maintaining entropy in RB assignments, and satisfying minimum-rate constraints. After the peak, the loss begins a consistent decline, falling below 2×10^4 by epoch 30 and stabilizing around 1.3×10^4 near epoch 120. The smooth tail indicates convergence and balanced gradient flow between the two task heads. This steady reduction demonstrates that the optimization successfully aligns the learned allocation patterns with those of the teacher model while maintaining training stability.

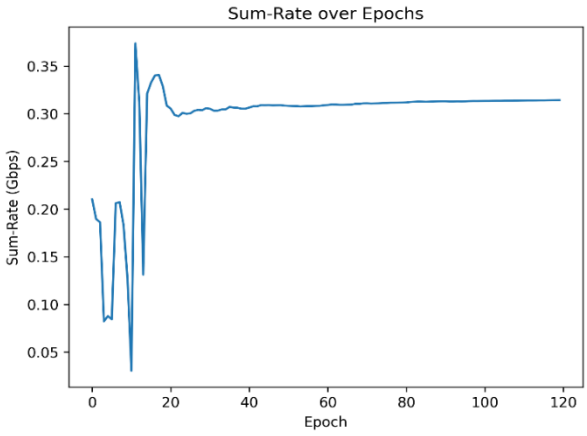

Fig. 7. Sum-Rate over Epochs.

Aggregate network throughput evolution, illustrated in Fig. 7, follows a pattern similar to the loss curve. The sum-rate initially oscillates sharply between 0.03 Gbps and 0.38 Gbps within the first 15 epochs, reflecting unstable association-power interactions before the loss terms are properly balanced. After epoch 20, the throughput stabilizes around 0.31–0.32 Gbps, maintaining this value for the rest of the training process. This behavior confirms that the model learns a stable resource-allocation strategy early in the joint training stage. The final throughput matches the evaluation result of approximately 0.32 Gbps, demonstrating that training convergence directly translates to consistent test-time performance. The brief overshoot near 0.38 Gbps during early epochs represents a temporary over-allocation to high-SINR users before fairness constraints take effect, showing the model's gradual balancing between efficiency and equity.

Table 2. Methods and Complexity

Method	**Computational Complexity**
Teacher (WMMSE)	O(U3)+O(B·RB2)
Baseline (RSSI + Equal Power)	O(U·B)
Proposed HetGNN	O(｜V｜+｜E｜)

Finally, Table 2 compares the computational complexity and runtime characteristics of the WMMSE teacher, the baseline heuristic, and the proposed HetGNN. While the WMMSE algorithm provides near-optimal results, its iterative nature results in high computational cost and latency, making it impractical for real-time 5G scheduling. The baseline method offers extremely low complexity but lacks interference awareness and power control precision. In contrast, the proposed HetGNN achieves a balanced trade-off: it maintains more than 85% of the teacher's throughput while performing inference in a single forward pass with linear complexity O(｜V｜+｜E｜). This enables decision-making within a few milliseconds, demonstrating the framework's suitability for real-time and large-scale network operation.

VI. Conclusion and Future Work

This section summarizes the main findings derived from the performance evaluation and training analysis of the proposed HetGNN model. The objective was to determine whether the network can jointly learn user association and antenna-level power allocation in a single inference step, while maintaining performance close to the WMMSE teacher under realistic 5G MIMO HetNet conditions. The experimental results confirm that the HetGNN successfully converges to a stable operating point, balancing spectral efficiency and learning stability. Two key indicators training loss and aggregate sum-rate highlight the model's effectiveness and its ability to approximate the teacher's behavior without iterative optimization.

The results demonstrate that the HetGNN successfully achieves this objective. Despite not exceeding the baseline or the WMMSE teacher in all quantitative metrics, the model consistently approximates the teacher's behavior with far lower computational cost and latency. The stable convergence of both training loss and aggregate sum-rate confirms that the network effectively learns to balance user distribution and power allocation within a single inference step. From a system-level perspective, the HetGNN reproduces more than 85% of the teacher's optimal throughput while maintaining acceptable fairness and substantially faster inference time. This trade-off highlights the practical strength of graph-based learning approaches for real-time resource coordination.

Overall, the findings validate that a heterogeneous graph neural network can serve as a scalable and efficient alternative to classical optimization methods, providing near-optimal performance with greatly reduced complexity. The proposed approach therefore represents a promising foundation for intelligent, low-latency resource management in future 5G and beyond networks.

Although the proposed graph includes antenna nodes, the optimization variables and evaluation metrics are defined at the BS/RB level, rather than at the true per-antenna level. In addition, the channel model is based on an equivalent scalar gain representation and does not explicitly model adaptive beamforming vectors or fully general MIMO precoding. The reported evaluation is also limited to a simplified 5 BS Uma inspired scenario with a spatial hold out split. For this reason, broader claims regarding generalization across different layouts, traffic regimes, or deployment conditions would require further experimental validation. Finally, the specific contribution of the heterogeneous graph construction has not yet been fully isolated through direct comparison with homogeneous GNN models or simpler BS only graph representations. These considerations do not diminish the value of the present proof of concept study, but they do define the appropriate boundaries within which its conclusions should be understood.

Future research can extend the present framework in several directions to further enhance its accuracy and adaptability. One promising avenue involves integrating reinforcement learning or meta-learning mechanisms to enable the HetGNN to refine its power allocation and association strategies dynamically based on real-time network feedback. Such an adaptive approach could allow the model to learn from changing channel conditions and user mobility patterns, thereby maintaining near-optimal performance even under non-stationary network environments. Another potential direction lies in exploring multi-agent and federated graph learning paradigms, where individual BSs or network clusters collaboratively train local sub-models while preserving data privacy and reducing communication overhead. Combining this with transfer learning could allow the HetGNN to generalize across different deployment scenarios or frequency bands without complete retraining. Additionally, incorporating explicit energy efficiency and quality-of-service objectives into the loss function would make the framework more suitable for large-scale green 5G deployments and future 6G architectures emphasizing sustainable and intelligent network operation.

REFERENCES

[1] I. A. Bartsiokas, P. K. Gkonis, D. I. Kaklamani and I. S. Venieris, "ML-Based Radio Resource Management in 5G and Beyond Networks: A Survey," in IEEE Access, vol. 10, pp. 83507-83528, 2022.

[2] K. M. Ayoub, et al, "Resource allocation schemes for 5G network: A systematic review", Sensors 21.19 (2021): 6588.

[3] A. Mamane, M. Fattah, M. E. Ghazi, M. E. Bekkali, Y. Balboul and S. Mazer, "Scheduling Algorithms for 5G Networks and Beyond: Classification and Survey," in IEEE Access, vol. 10, pp. 51643-51661, 2022.

[4] C. Robin, and R. Akl, "Massive MIMO systems for 5G and beyond networks—overview, recent trends, challenges, and future research direction", Sensors 20.10 (2020): 2753.

[5] 3GPP TR 38.901, "Study on channel model for frequencies from 0.5 to 100 GHz", 3rd Generation Partnership Project, Release 15, 2020.

[6] H. Yang et al., "Knowledge-Driven Resource Allocation for Wireless Networks: A WMMSE Unrolled Graph Neural Network Approach", in IEEE Internet of Things Journal, vol. 11, no. 10, pp. 18902-18916, 15 May15, 2024

[7] M. Ali, Q. Rabbani, M. Naeem, S. Qaisar and F. Qamar, "Joint User Association, Power Allocation, and Throughput Maximization in 5G H-CRAN Networks", in IEEE Transactions on Vehicular Technology, vol. 66, no. 10, pp. 9254-9262, Oct. 2017.

[8] S. Jamshidiha, V. Pourahmadi and A. Mohammadi, "A Traffic-Aware Graph Neural Network for User Association in Cellular Networks", in IEEE Transactions on Mobile Computing, vol. 24, no. 8, pp. 6858-6869, Aug. 2025

[9] Shannon, Claude E, "A mathematical theory of communication", ACM SIGMOBILE mobile computing and communications review 5.1 (2001): 3-55

[10] Q. Shi, M. Razaviyayn, Z. -Q. Luo and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011, pp. 3060-3063

[11] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for wireless resource management", IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Sapporo, Japan, 2017, pp. 1-6

[12] A. Chowdhury, G. Verma, C. Rao, A. Swami and S. Segarra, "Unfolding WMMSE Using Graph Neural Networks for Efficient Power Allocation", in IEEE Transactions on Wireless Communications, vol. 20, no. 9, pp. 6004-6017, Sept. 2021

[13] K. Shen and W. Yu, "Fractional Programming for Communication Systems—Part I: Power Control and Beamforming", in IEEE Transactions on Signal Processing, vol. 66, no. 10, pp. 2616-2630, 15 May15, 2018

[14] O.S. Falade, "DeepMIMO: A Generic Deep Learning Dataset for Millimeter Wave and Massive Mimo Applications to Vehicular Communications," Available at SSRN 4383745 (2023).

[15] 3rd Generation Partnership Project (3GPP). (n.d.). 3GPP References Search. Retrieved from 3GPP website: https://portal.3gpp.org/3gppreferences/SearchReferences.aspx