Evaluating Defensive Influence in Multi-Agent Systems Using Graph Attention Networks

Gregory Everett[1], Ryan J. Beal[2], Tim Matthews[2], Timothy J. Norman[1], Sarvapali D. Ramchurn[1]
[1]School of Electronics and Computer Science, University of Southampton, United Kingdom
[2]Sentient Sports, Southampton, United Kingdom
gae1g17@soton.ac.uk, ryan.beal@sentientsports.com, tim.matthews@sentientsports.com,
t.j.norman@soton.ac.uk, sdr1@soton.ac.uk

***Abstract*—Evaluating individual contributions from team members is a critical challenge across many domains, such as security and team sports. While progress has been made in valuing contributions, such as target defence in security or on-ball performance in football (soccer), many aspects of performance, such as off-ball football actions, remain difficult to quantify. We introduce *GAPP*, a Graph Attention Network model that predicts football pass reception probabilities and provides interpretable insights into off-ball defending. Using attention mechanisms, GAPP captures player interactions and introduces two new metrics to quantify defender contributions. We tested GAPP on 306 English Premier League matches, and showed it reduces binary cross-entropy loss by 6.4 percent compared to multiple baselines for pass reception prediction, while offering unique insights for off-ball defender evaluation for coaches, scouts and teams. This work shows the potential of graph attention networks for analysing complex multi-agent systems like football.**

***Index Terms*—Graph Attention Networks, Sports Analytics, Football, Applied Machine Learning**

I. Introduction

Evaluating individual agent performance within a team is critical in many real-world domains, as it can provide actionable insights for improvement and optimise team strategies. In fields like security [1] and sports analytics [2], [3], performance is often assessed through measurable actions, such as stopping a security breach or completing a high-value pass. However, many contributions are indirect and harder to quantify, such as an agent's positioning influencing an attacker's target choice in security or a defender's off-ball positioning affecting pass reception in team sports. Traditional metrics often overlook these subtle impacts, highlighting the need for methods that evaluate indirect contributions to gain deeper insights into performance and improve decision-making.

In this paper, we focus on evaluating off-ball performance in Association football (soccer), which has rich availability of real-world spatiotemporal datasets. In football, players position themselves to prevent dangerous opposition attackers from receiving the ball, yet these off-ball contributions are challenging to quantify due to the lack of direct links to outcomes like goals, unlike on-ball events such as passes or shots, which are easily measurable. Players spend the majority of a match ($\sim$95%) off the ball, making off-ball performance a critical yet underexplored aspect of the game.

The availability of spatiotemporal datasets, such as event data (e.g., passes, shots) and tracking data (e.g., player locations), have advanced football analytics, enabling machine and deep learning models to evaluate player performance, primarily focusing on on-ball actions like attacking plays [2], [4] or defensive actions [3]. However, assessing the impact of off-ball defensive positioning remains challenging due to its indirect influence on play outcomes in a dynamic environment. While some models predict future ball locations based on player positioning [4], they do not quantify the specific influence of individual defenders on these outcomes.

Against this background, we introduce *GAPP* (Graph Attention for Pass Probabilities), a novel Graph Attention Network [5] (GAT)-based model that predicts the probability of attacking players receiving the ball while providing interpretable insights into off-ball defensive contributions. By representing players and the ball as a dynamic graph, GAPP leverages GATs' attention mechanism to identify key relationships between players. The GAPP model not only outperforms baselines in pass reception prediction but is also used to provide two novel, attention-based metrics for evaluating off-ball individual defensive contributions.[1]

Thus, this paper presents the following novel contributions:

- We introduce GAPP, a novel graph attention-based model for predicting pass reception probabilities in football.
- GAPP achieves state-of-the-art performance for football pass prediction, outperforming multiple baselines, including a traditional graph neural network (GNN), with a $\sim$6.4% $\pm$ 1.5% reduction in binary cross-entropy loss.
- We introduce two novel attention-based metrics to quantify defender influence (DI) on attackers and defensive performance (DP) in a real-world multi-agent system.
- Using real-world event and tracking data from 306 English Premier League (EPL) football matches, we show how GAPP provides explainable insights into defender performance, including their impact on stopping dangerous attackers.

Gregory Everett was supported by Sentient Sports and Sarvapali Ramchurn was supported by the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1) and Responsible AI UK (EP/Y009800/1).

[1]A public repository containing the code and GAPP model presented in this paper, as well as an appendix for this paper, is available here: https://github.com/GregSoton/EvaluatingDefensiveInfluenceUsingGATs.

In Section II, we review related work. Section III formalises the ball reception prediction problem, and Section IV introduces our GAPP model. Section V introduces our defensive metrics. Section VI presents empirical evaluations, followed by a case study on the EPL in Section VII. Section VIII discusses the results, and Section IX concludes.

II. Related Work

In this section, we review existing research for player evaluation, graph-based models, and attention mechanisms for interpretability.

A. Player Evaluation in Football

Player evaluation in football is challenging due to the sport's fluid and unpredictable nature, but spatiotemporal datasets have enabled more advanced player evaluation metrics beyond goals and assists. For instance, Expected Goals (xG) models [6] estimate the probability that a given shot will result in a goal by leveraging spatiotemporal features such as the positions of the ball and defenders. Building on this, models have also been proposed to assess actions other than shots. For example, Expected Threat (xT) [7] quantifies the value of on-ball attacking actions, such as passes, by estimating the likelihood of scoring from a particular ball location, based on past event data. Similarly, VAEP (Valuing Actions by Estimating Probabilities) [2] calculates changes in scoring and conceding probabilities after on-ball actions using a CatBoost model [8]. However, these metrics ignore the impact of off-ball player positioning. Spatiotemporal tracking data has enabled metrics that incorporate broader match context [4], [9]. Fernandez et al. [4], [10] introduce SoccerMap, a convolutional neural network (CNN) that uses tracking data to predict pass success and goal probability. Robberechts et al. [9] build on this model by identifying unpredictable but successful passes to quantify player creativity in football. Rahimian et al. [11] use a possession model to analyse on-ball actions and leverage reinforcement learning to derive the optimal policy in football situations, allowing comparison between actual human actions and a suggested optimal action.

These metrics evaluate on-ball attacking actions instead of defensive contributions. Merhej et al. [3] address this by evaluating defensive actions (e.g., tackles, interceptions) using a Multi-Layer Perceptron and event data to predict scoring probabilities had the action not occurred. However, this is limited to on-ball actions, overlooking off-ball defending, which is more common as players spend $\sim$95% of a match off the ball. While some studies assess overall team defence [12], [13], to our knowledge, no existing model evaluates the influence of individual off-ball defenders on attackers.

B. Graph-Based Approaches in Football

Graph-based learning methods have successfully modelled football due to the dynamic relational properties between teams. Previous work has used GNNs to predict player movement [14], [15]. For example, Everett et al. [14] use graph convolutional networks (GCNs) to estimate tracking data from on-ball events. Similarly, Yeh et al. [16] apply GNNs to predict player trajectories in football and basketball.

Graph-based models have also been applied to football outcome predictions. Rahimian et al. [17] use temporal GNNs to predict actions (e.g., pass or shot success). Wang et al. [18] apply a GAT to predict corner outcomes, such as receiver and shot predictions. However, their work focuses solely on corners, while our model generalises to all phases of play. Stöckl et al. [19] use a GCN to predict attacker scoring probabilities and assess team-level defence. However, these methods do not evaluate individual defensive contributions. Our model uniquely leverages attention weights to create interpretable metrics for off-ball defensive player contributions.

C. Attention Mechanisms for Interpretability

Attention mechanisms enable machine learning models to focus on the most relevant input data, improving predictive performance and interpretability [20]. This is widely used in natural language processing [21], computer vision [22] and graph-based modelling [23], [24]. We focus on graph-based modelling in this work.

GNNs are effective for prediction tasks with relational data, such as social networks [25] and sports teams [14], [18]. Traditional GNNs, such as GCNs, aggregate information equally from connected nodes through message passing [26], which limits their ability to capture node importance. GATs overcome this limitation by using attention mechanisms to assign context-dependent, learnable weights during node aggregation [5], [27], improving model performance and interpretability by identifying the most influential nodes [27], [28]. This makes GATs a significant advancement in modelling complex, dynamic systems like football.

Other recent advances in deep graph learning have made progress in knowledge graph reasoning, interpretability, and model robustness. Zhang et al. [29] propose adaptive propagation paths for GNNs that identify semantically related entities, improving the interpretability and efficiency of reasoning in complex relational domains. Zhou et al. [30] present the Robust Graph Information Bottleneck (RGIB) method to improve the robustness of link prediction in GNNs when there is bilateral edge noise, a common issue occurring from noisy data in real-world multi-agent domains. Additionally, Zhou et al. [31] propose scalable, query-dependent subgraph extraction to improve link prediction efficiency on large-scale graphs. Our work aims to further understand complex relational systems by extracting interpretable defensive influence metrics from a GAT, tailored for dynamic multi-agent environments.

While GATs offer benefits, extracting interpretability can be challenging, as high attention weights do not always clearly indicate positive or negative impacts, and some studies suggest they may not reliably reflect true importance in certain settings [32]. Inspired by node masking methods to extract node influence in GNNs [33], we propose a framework to extract interpretable influence metrics from our GAT model. To our knowledge, our approach to evaluating individual contributions in a complex multi-agent system is novel.

III. Predicting Ball Reception

In this section, we model a game scenario and formalise the problem of predicting football on-ball event (e.g., pass, dribble) outcomes. A game is represented as a time series of T events, $E = [e_1, ..., e_T]$, where each event $e_t \in E$ corresponds to an on-ball action. This time series is non-uniform, with varying intervals between events.

For each event $e_t \in E$, there is a set of N players, $A = \{a_1, ..., a_N\}$, who are involved in the game at that event. Each player belongs to one of two teams, determined by the function team : $A \to \{1, 2\}$. While players remain on the same team throughout the game, the roles of the teams (attacking vs. defending) can switch dynamically over the sequence of events E depending on which team has ball possession. At each event e_t, the player performing the on-ball action is denoted as a_n^t, where $a_n^t \in A$. The team of a_n^t, denoted as team(a_n^t), forms the attacking team A^O at that event, i.e., $A^O = \{a \in A : \text{team}(a) = \text{team}(a_n^t)\}$. The remaining players, belonging to the opposing team, form the defending team $A^D = A \setminus A^O$.

Our goal is to model the probability of a player $a \in A$ receiving the ball at the next event. Let e_t represent the current event and e_{t+1} the next event in the sequence E. At each event, we aim to predict the probability of each player $a_n \in A$ being the on-ball player at e_{t+1}, defined as $\Pr(a_n^{t+1}|e_t)$, where a_n^{t+1} indicates that player n possesses the ball at the next event. Accurately estimating this probability helps evaluate the likelihood of each player scoring and the value of the defending team's positioning (Section V).

To model the probability of a player receiving the ball at the next event, we frame this as a graph learning problem. Each game event e_t is represented as a graph $G_t = (V_t, \xi_t)$, where V_t is the set of nodes and ξ_t the set of edges. The nodes V_t consist of the players $A = \{a_1, ..., a_N\}$ and the ball, denoted as a special node b, such that $V_t = A \cup \{b\}$. Edges capture relationships between players. Each node $v \in V_t$ has a feature vector $\mathbf{x}_v \in \mathbb{R}^d$, and each edge $(u, v) \in \xi_t$, where u and v are nodes, has a feature vector $\mathbf{y}_{u,v} \in \mathbb{R}^k$, where d and k are the dimensions of the node and edge features.

At each event e_t, predicting the next on-ball player a_n^{t+1} is framed as a node prediction task. We learn a function $f : G_t \to [0, 1]$ that maps the graph to the probability of a specific player a_n receiving the ball at the next event: $\Pr(a_n^{t+1}|e_t) = f(G_t)$. The GAPP model represents the function f and is introduced in the next section.

IV. GAPP Model For Ball Receiver Prediction

In this section, we outline the feature engineering (Section IV-A), model architecture (Section IV-B), and training process (Section IV-C) used for our GAPP model.

A. Feature Engineering

The node and edge features of our graph G_t are derived from football tracking data, which provides the coordinate locations, velocities and accelerations of each player and the ball on the pitch at a given timestep. However, these raw data lack important contextual information, such as the relationships between players and their positions relative to the goals. To address this, we apply feature engineering to extract features for our model. Specifically, for each graph G_t, we compute the following node and edge features:

- **Node Features:** Location (x, y), velocity (x, y), acceleration (x, y), distance to defending goal (x, y), distance to attacking goal (x, y), Euclidean distance and angle (radians) to the ball, binary indicators for whether the node is on the home team, attacking, and on the ball.
- **Edge Features:** Distance between nodes (x, y), Euclidean distance between nodes, edge angle (radians), difference in node angles to the ball, binary indicator for nodes being on the same team.

These features are selected to provide a comprehensive tactical representation of each event. For example, the velocity and acceleration features contribute to capturing aspects of the immediate temporal context and player intention. For the ball node, numeric ball-related features are set to 0, and binary features are set to 1. For the same-team binary indicator edge feature, where one node is the ball, this is 1 if the other node's team is attacking. Ball-related features are included to provide the model with easy access to this information during message passing. The orientation of play is consistent across all timesteps, with the attacking team always moving from left to right. For this work, the graph is fully connected and directed, with edges in both directions between every pair of nodes. A fully connected, directed graph enables the model to assess the tactical significance of interactions between all players, allowing the attention mechanism to learn and highlight the most important relationships in a data-driven manner. We define an adjacency matrix $\mathcal{A} \in \{0, 1\}^{|V_t| \times |V_t|}$ where $|V_t|$ is the number of nodes in graph G_t, such that $\forall u, v \in V_t : \mathcal{A}_{uv} = 1$, signifying full connectivity.

B. GAPP Model Architecture

Our GAPP model represents the function f, which predicts reception probability $\Pr(a_n^{t+1}|e_t)$ using a graph representation G_t and GATs, which dynamically learn the importance of node connections in graph-structured data. In team sports, interactions between players and the ball are complex and vary in significance. Unlike traditional GNNs, which assign equal attention to all node edges, GATs assign attention weights to edges, capturing the varying importance of player relationships for event prediction. The model's step-by-step architecture is given below and visualised in Figure 1.

1) Step 1: Encoding Layer: The input to the GAPP model consists of node features and edge features, represented as tensors $\mathbf{X_t} \in \mathbb{R}^{B \times |V_t| \times d}$ and $\mathbf{Y_t} \in \mathbb{R}^{B \times |\xi_t| \times k}$, where B is the batch size. The GAPP model uses dense encoding layers to project node and edge features into latent spaces, allowing the model to learn meaningful representations of these nodes and edges. The encoding outputs are latent representations $\mathbf{H_V} \in \mathbb{R}^{B \times |V_t| \times H_V}$ and $\mathbf{H}_\xi \in \mathbb{R}^{B \times |\xi_t| \times H_\xi}$ where H_V and H_ξ are the hidden layer sizes of the node and edge encodings, respectively. In this work, $B = 64$, $H_V = 32$ and $H_\xi = 16$.

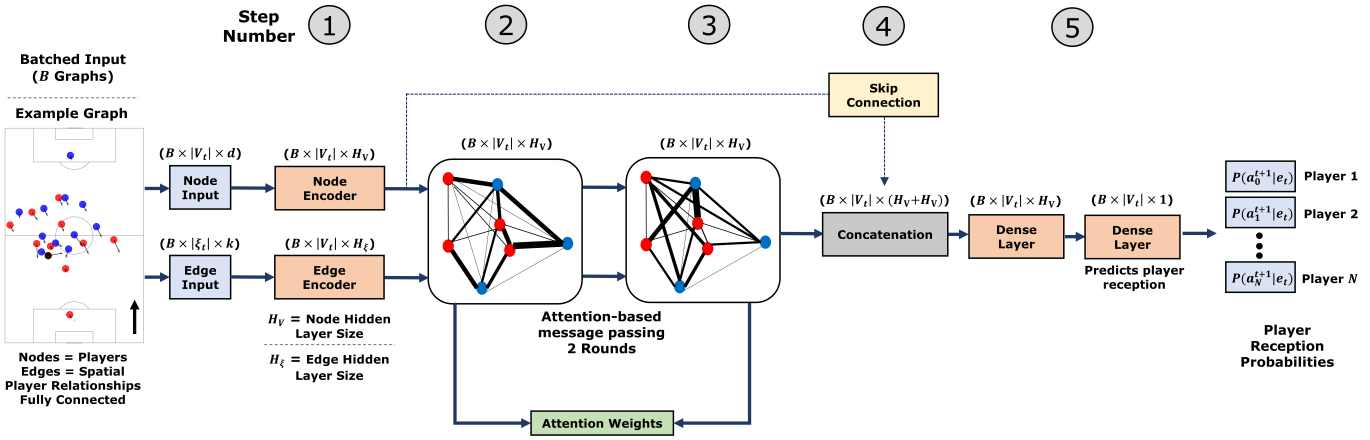

Fig. 1. GAPP model architecture. All predictions are made using player feature information described in Section IV-A, with batch size B, number of nodes $|V_t|$, number of edges $|\xi_t|$, number of node features d, and number of edge features k.

2) Step 2: GAT Layer 1: The second step of the GAPP model is the first GAT layer, which performs attention-based message passing [5]. This layer takes node encodings $\mathbf{H_V}$, edge encodings $\mathbf{H}_\xi$ and the adjacency matrix $\mathcal{A}$ as input. The GAT computes attention weights α_{uv} for each edge $(u, v) \in \xi_t$, reflecting the relevance of node v and the edge connecting it to u. These attention weights are derived from raw scores β_{uv} based on node and edge features and normalised with a softmax over u's neighbours so the weights sum to 1. Each node updates its representation by aggregating the features of its neighbouring nodes, weighted by their respective attention scores. This allows the model to focus on the most relevant neighbours during message passing. Note that both GAT layers include self-loops and use 16 attention heads in this paper, meaning that α_{uv} is a vector of 16 values.

The output of this layer is a new set of node embeddings $\mathbf{H_V^{'}} \in \mathbb{R}^{B\times|V_t|\times H_V}$, where the attention mechanism adaptively aggregates graph information. In football, this models the influence of players on each other towards ball reception. A ReLU activation function is then applied to the output embeddings $\mathbf{H_V^{'}}$, followed by a dropout layer with a zeroing probability of 0.1, resulting in updated embeddings $\mathbf{H_V^{'}}$.

3) Step 3: GAT Layer 2: The third step of the GAPP model is a second GAT layer, which captures higher-order relationships between nodes in the graph through another round of attention-based message passing. This layer takes the updated node embeddings $\mathbf{H_V^{'}}$ from the first GAT layer, the edge encodings $\mathbf{H}_\xi$ and the adjacency matrix $\mathcal{A}$ as input. Message passing is performed as in the previous layer, and the output is a new set of node embeddings $\mathbf{H_V^{''}} \in \mathbb{R}^{B\times|V_t|\times H_V}$. As in step 2, a ReLU activation function and dropout layer are then applied to the new node embeddings. In the context of football, the second GAT layer enables the model to capture more complex relationships between players in the graph.

4) Step 4: Skip Connection and Concatenation: The original encoded node features $\mathbf{H_V}$ are concatenated with the updated node embeddings $\mathbf{H_V^{''}}$ through a skip connection. This concatenation enables the model to retain original node information while including higher-order node (player) relationships learned by the GAT layers. The resulting concatenated embeddings $\mathbf{H_V^{'''}} \in \mathbb{R}^{B\times|V_t|\times(H_V+H_V)}$ are then used as input for the final dense layers of the model.

5) Step 5: Output Dense Layers: The concatenated embeddings $\mathbf{H_V^{'''}}$ pass through a dense layer with ReLU, resulting in a final latent representation for each node, $\mathbf{H_V^{*}} \in \mathbb{R}^{B\times|V_t|\times H_V}$. This representation is then passed through another dense layer followed by a softmax activation function, resulting in the final node predictions with shape $(B \times |V_t| \times 1)$. These predictions represent the probability, ranging from 0 to 1, of each player receiving the ball at the next timestep.

C. Model Training

We use football tracking data, which records all player locations during every on-ball event from 306 EPL matches in the 2023/24 season (see Section VI-A), to extract our target labels for the GAPP model, where each player is assigned a label of 1 if they are the next on-ball player, and 0 otherwise, resulting in one positive label at each timestep. For model training, we focus on attacking player receptions by applying an attacking player mask to the loss function, ensuring that updates are made only for attacking players. The model is trained using a binary cross-entropy (BCE) loss function for 200 epochs with a batch size of 64. The Adam optimiser [34] is used with an initial learning rate of 0.003. The GAPP model hyperparameters and architecture were determined through trial and error with no formal hyperparameter tuning.

V. Valuing Defensive Positioning

The key goal of the GAPP model is to provide insights into how defender positioning impacts attacker reception probabilities, using the GAT attention mechanisms to highlight the most relevant nodes and edges towards predictions. This section explains how attention weights are used to derive a novel defensive influence and defensive performance metric for evaluating off-ball defensive contributions.

A. Extracting Defender Influence

A defender's attention weight shows their influence on an attacker's reception probability. However, in its raw form, the attention weight only indicates the relative importance of the defender without specifying whether their impact is positive or negative. To address this, we compute a more interpretable defender influence (DI) metric that quantifies the effect of a defender on an attacker's reception probability.

Let a node v represent a defender $a_d \in A^D$ and a node u represent an attacker $a_o \in A^O$. The attention weights α_{uv} quantify the attention assigned to defender v for attacker u, computed separately for each GAT layer and across all attention heads. The probability of attacker a_o receiving the ball, $\Pr(a_u^{t+1}|e_t)$, is predicted using the trained GAPP model f. To measure defender influence, we recompute the reception probability with α_{uv} masked to 0 in both GAT layers: $\Pr(a_u^{t+1} \mid e_t, \alpha_{uv} = 0)$, i.e. the prediction without defender v's attention. This masking allows the GAPP model to predict the reception probability of the attacker without considering the influence of defender v. This approach is inspired by explanation methods in the literature that assess prediction changes by limiting node features [33], but we uniquely apply it to quantify defensive contributions in a multi-agent system.

Defenders with higher attention weights are expected to have greater impacts on recomputed probabilities than defenders with lower weights (see Section VI-E). Using these probabilities, we define a new defensive influence (DI) metric for defender v on attacker u:

$$\mathrm{DI}_{uv} = \Pr(a_u^{t+1}|e_t, \alpha_{uv} = 0) - \Pr(a_u^{t+1}|e_t) \tag{1}$$

The DI metric quantifies a defender's impact on an attacker's ball reception probability. A positive DI_{uv} indicates the defender reduces the attacker's probability by that amount.

B. Computing Defender Performance

To evaluate the overall performance of a defender v, we first model the attacking threat of an attacker u. To measure attacking threat, we use the Expected Threat (xT) metric introduced by Singh [7], which divides the football pitch into a grid of zones (16×12 for this work) and estimates the probability of scoring within the next 5 on-ball events from each zone based on historical data. The threat of an attacker u, denoted as xT_u, is determined by identifying the zone they occupy and assigning the corresponding xT value.

We combined each attacker's threat (xT_u) with the defender's influence on them (DI_{uv}) to calculate defender v's overall positional performance. This is captured by our novel defender performance (DP) metric for a defender v defined as:

$$\mathrm{DP}_v = \sum_{u \in A^O} \mathrm{DI}_{uv} \cdot \mathrm{xT}_u \tag{2}$$

This metric provides a novel evaluation of defender v's positioning by aggregating their influence on all attackers weighted by each attacker's scoring threat, offering a measure of a player's off-ball defensive contribution at an event e_t.

VI. Empirical Evaluation

In this section, we present our dataset, baseline methods, and empirical evaluation of the GAPP model and DP metric.

A. Datasets

Our GAPP model is trained and evaluated on 306 EPL games from the 2023/24 season. This data was supplied to us by Gradient Sports. The dataset includes event data (e.g., passes, shots) and tracking data (player positions and velocities). This is a gold standard industry dataset, enabling a rigorous evaluation of our model. We use tracking frames aligned with on-ball events to construct our feature set and target variables, resulting in 359,040 events. We evaluate our model using five-fold cross-validation, splitting the games into $\sim$80% for training and $\sim$20% for testing (245/61) in each fold, ensuring that each game appears in the test set exactly once.

B. Baselines

We evaluate GAPP's performance against several baselines:

- **Distance** - a baseline where reception probability is inversely proportional to a player's distance from the ball.
- **Spearman** - a physics-based model adapted from [35] to predict reception probabilities instead of pass locations.
- **Dauxais** - a random forest using distance-based features [36], using the majority of features from the original paper, extended with our node and edge features.
- **XGBoost** - an XGBoost model [37] using the same node and edge features as GAPP. XGBoost is shown to be an effective predictor of pass reception in prior studies [9].
- **Graph Neural Network** - a GNN with the same architecture as GAPP but replacing GAT with SAGEConv [26] layers without attention.

These baselines provide a diverse range of approaches, from simple distance-based heuristics to advanced machine learning and graph-based methods, enabling a rigorous evaluation of our model against current pass reception prediction models.

C. Experiment 1: GAPP Model Accuracy

We evaluate the predictive performance of GAPP and baselines towards pass reception in Table I. We use various metrics, averaged across all attacker nodes, to assess the calibration of the model's predicted pass reception probabilities.

TABLE I
PREDICTIVE PERFORMANCE OF MODELS FOR PASS RECEPTION AVERAGED ACROSS FIVE FOLDS, ALONG WITH 95% CONFIDENCE INTERVALS. BOLD RESULTS INDICATE THE BEST PERFORMANCE.

Models	Test BCE	AUC Score	F1 Score
Distance	0.304 ± 0.005	0.655 ± 0.005	0.866 ± 0.001
Spearman	0.298 ± 0.001	0.690 ± 0.003	0.866 ± 0.000
Dauxais	0.286 ± 0.001	0.698 ± 0.003	0.867 ± 0.001
XGBoost	0.264 ± 0.001	0.760 ± 0.004	0.869 ± 0.001
GNN	0.233 ± 0.002	0.830 ± 0.005	0.897 ± 0.001
GAPP	**0.218 ± 0.003**	**0.855 ± 0.006**	**0.910 ± 0.002**

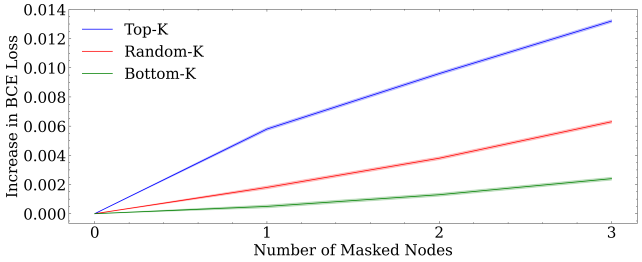

Fig. 2. Fidelity evaluation of the GAPP model, illustrating the increase in BCE loss when masking various numbers of nodes for each masking strategy. Shaded 95% confidence boundaries are included but are minimal.

We find that GAPP achieves the best performance across all metrics, with a ∼6.4% reduction in BCE loss compared to the GNN model with the same architecture, highlighting the benefit of the attention mechanism for learning the importance of player relationships and interactions towards pass reception probability. Additionally, GAPP outperforms the XGBoost model with a ∼17.4% reduction in BCE loss, highlighting the advantages of using a graph-based approach to model the spatial and tactical context of game situations.

D. Experiment 2: Testing Model Fidelity

We evaluate the effectiveness of GAPP's attention mechanism by testing its impact on pass reception predictions. We test this using various metrics introduced in the literature for evaluating attention-based models. Specifically, we use the fidelity metric [38], a well-established metric for evaluating GNN models, which measures the deterioration in model performance (BCE Loss) when specific defender edge attentions are masked to 0. We compare three masking strategies:

- Top-K - Masking the K defenders receiving the highest average attention from the attacker (averaged across all heads for the attacker-defender edge).
- Random-K - Masking a random selection of K defenders.
- Bottom-K - Masking the K defenders receiving the lowest average attention from the attacker.

We conduct this analysis for K = 1, 2, and 3, in Figure 2.

The results show that masking the Top-K defenders leads to the largest increase in BCE Loss, with a 5.8×10^{-3} increase for K=1 compared to 1.8×10^{-3} for Random-K and 5.0×10^{-4} for Bottom-K. This trend continues for K=2 and K=3, demonstrating that high-attention defenders contribute significantly to the model's predictions, while low-attention defenders have minimal impact.

E. Experiment 3: Testing Model Faithfulness

We perform additional tests on the model's attention mechanism by testing GAPP's faithfulness, which assesses how well the attention mechanism aligns with the model's reasoning process [39], [40]. Faithfulness is typically tested by manipulating inputs and observing prediction changes. We test faithfulness by setting a defender's attention to 0 and measuring the change in an attacker's reception probability. Linear regression shows that a 0.1-unit increase in defender attention corresponds to a ∼6.9% increase in the relative percentage change of an attacker's reception probability when the defender's attention is removed (coefficient = 69.3, $p < 0.01$). Figure 3 presents a scatter plot based on a stratified sample of the data illustrating this observed relationship.

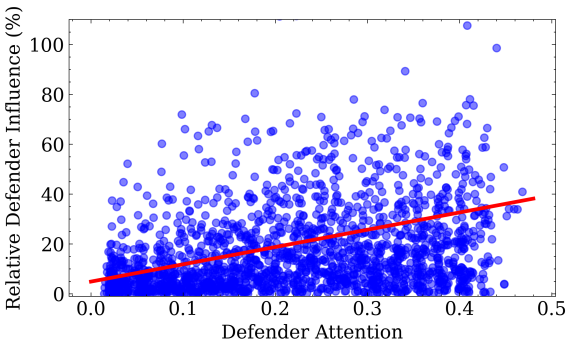

Fig. 3. Assessing GAPP faithfulness using the correlation between defender attention and relative defender influence.

This finding suggests that the model's attention mechanism is closely aligned with changes in its predictions. This alignment further supports the idea that the attention scores capture key features influencing the model's pass reception decisions.

F. Experiment 4: Impact of Defender Performance on Defensive Actions

In this experiment, we evaluate the relationship between the DP metric and the probability of a defender performing a defensive action. Using logistic regression, we model the likelihood that a defender performs a defensive action within the next three on-ball events given their relative performance. Relative performance is calculated as the difference between a defender's DP and their team's average DP at the same event.

The analysis shows a significant positive relationship (coefficient = 0.762, $p < 0.01$), where a 0.1-unit increase in relative performance corresponds to a 7.9% increase in the probability of performing an on-ball action. This suggests that higher relative performance improves the likelihood of timely defensive actions. These results may vary for specific teams and attacking opponents. The findings also suggest that defenders actively disrupting attacker receptions are more likely to perform an action, which may indicate that proximity is influencing their impact. The standard deviation of relative performance is ∼0.1 (mean ≈ 0) in our data, showing the potential for players to improve their likelihood of defensive action with improved off-ball defensive positioning.

VII. Model Application to English Premier League

In this section, we present our model's results on 2023/24 EPL data, analysing GAPP model performance and comparing defenders using the DP metric across varying contexts.

A. Model Accuracy Across Pitch Zones

We analyse how context affects GAPP's prediction accuracy. Specifically, for our EPL dataset, we test the impact of ball location on model predictive accuracy in Figure 4.

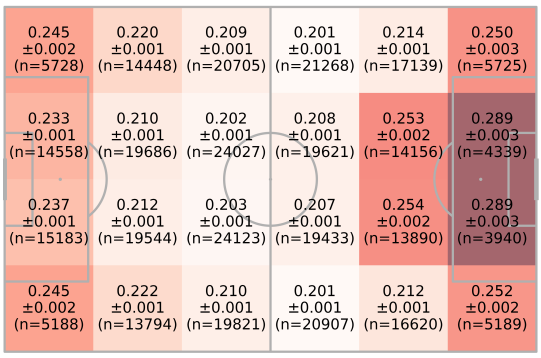

Fig. 4. BCE Loss of the GAPP model across our dataset for varying ball zones. Teams attack from left to right.

The model predicts passes more accurately in midfield and on the wings, suggesting that the structure of play in these areas makes actions easier to predict. Performance declines near the goal, where player positions are more congested, reflecting expert intuition that build-up play is more predictable than final attacking phases.

B. Visualising Off-Ball Defensive Contribution

We present visualisations to analyse the DI and DP metrics. These visuals translate our deep learning results and novel defensive metrics into an explainable framework for scenario analysis. Figure 5 shows a real-world example, highlighting Rico Lewis's influence (DI) on opposing attackers as predicted by the GAPP model.

This figure shows Rico Lewis's highest DI on attackers whose passing lanes are being blocked, likely increasing the chance of a backwards pass. Rico Lewis shows minimal influence for attackers positioned further away, where small values may reflect model noise. This visualisation illustrates how GAPP captures defensive spatial dynamics. Figure 6 shows the DP metric for each defender in a specific scenario.

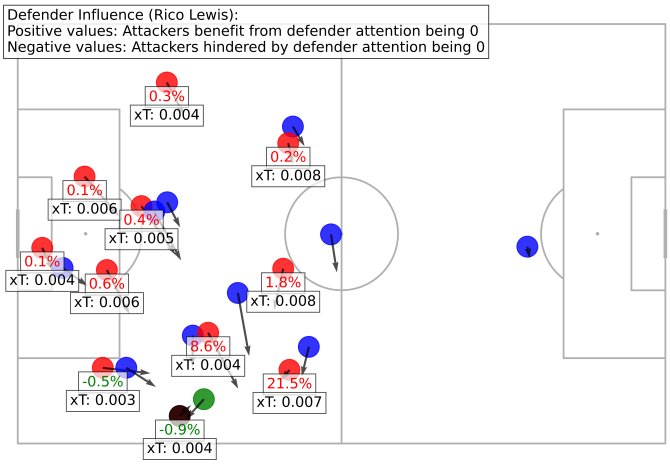

Fig. 5. An example visualisation showing Rico Lewis's (green defender) DI on each red team attacker, with xT values representing their threat. The red team attacks left to right.

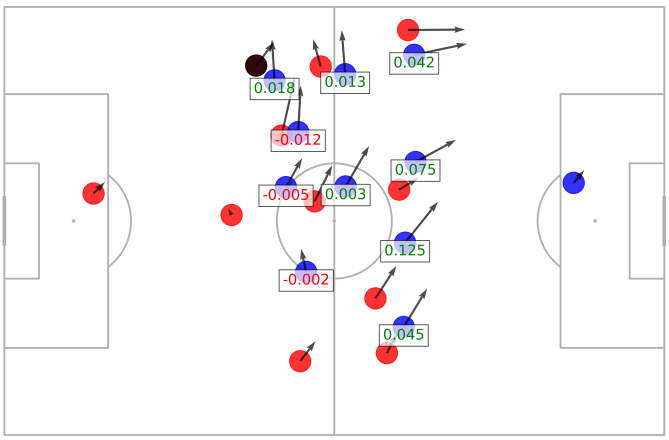

Fig. 6. An example visualisation showing the DP of each defender (blue). The shaded attacker is the ball carrier. Red team attacks left to right.

The DP metric shows that defenders marking high-threat attackers tend to have higher scores, reflecting the intuition that proximity to dangerous attackers is a critical factor in defending. These visualisations can analyse matches event-by-event or focus on key moments, offering a comprehensive tool to coaches and teams for evaluating defensive performance.

C. Rating Players Using Defender Performance

We compared the league's top center backs using their total DP metric for the season, analysing 99 center backs from our EPL dataset. To address potential biases, such as each team's number of defensive events or varying match counts in our dataset (which excludes some EPL games), we normalise the DP metric by the defenders team's number of defensive events in the dataset. The results are shown in Table II.

Interestingly, top defenders based on the DP metric mostly come from EPL clubs with around average league performance, likely because defender performance is valued more when influencing high-threat attackers in dangerous areas. Top teams defend less in these areas, leading to lower DP rankings. Similar trends appear in metrics like tackles and interceptions, which are often led by non-top team players. Among the 99 center backs in our dataset, those with high DP scores also typically rank highly in tackles and interceptions. Analysing defensive partnerships also provides valuable insights [41], as

TABLE II
RANKING EPL CENTER BACKS BASED ON DEFENDER PERFORMANCE (DP). IR AND TR ARE THEIR RANKINGS AMONG EPL CENTER BACKS FOR MOST INTERCEPTIONS AND TACKLES RESPECTIVELY.

Defender	Team	DP	Rank	IR	TR
F. Schär	Newcastle	0.0397	1	11	26
T. Gomes	Wolves	0.0395	2	27	21
J. Andersen	C. Palace	0.0394	3	7	6
M. Kilman	Wolves	0.0393	4	8	17
J. Tarkowski	Everton	0.0387	5	2	4
I. Zabarnyi	Bournemouth	0.0386	6	17	5
J. Branthwaite	Everton	0.0384	7	3	1

TABLE III
RANKING EPL CENTER BACK PARTNERSHIPS BASED ON COMBINED DEFENDER PERFORMANCE (DP). GC IS THE TEAM'S GOALS CONCEDED IN THE 2023/24 EPL SEASON.

Center Back Partnership	Team	DP	GC
C. Romero M. van de Ven	Tottenham	0.0496	61
J. Tarkowski J. Branthwaite	Everton	0.0444	51
F. Schär S. Botman	Newcastle	0.0431	62
Gabriel W. Saliba	Arsenal	0.0420	29

TABLE V
TOP 3 AND BOTTOM 3 EPL CENTER FORWARDS RANKED BY F. SCHÄR'S MEAN DI IN LIMITING THEIR BALL RECEPTIONS, BASED ON AT LEAST 200 EVENTS WHERE SCHÄR HAS THE HIGHEST ATTENTION.

Attacker	**Team**	**Mean DI**	**Minutes**	**xT/90**
		Top 3		
E. Adebayo	Luton	**0.0104**	166	0.48
R. Muniz	Fulham	**0.0072**	114	0.40
N. Jackson	Chelsea	**0.0068**	157	0.45
		Bottom 3		
D. Solanke	Bournemouth	**0.0057**	180	0.56
C. Archer	Sheffield United	**0.0054**	156	0.78
E. Haaland	Man City	**0.0042**	90	0.49

it allows managers to assess how well players work together. Table III ranks top center back pairings by combined DP, which is calculated only when both players are on the pitch together as the sole center backs. The DP is normalised by the team's total defensive events in our EPL dataset.

These results highlight defensive partnerships that coordinate well off the ball. We also include each partnership's team and average goals conceded, where for all teams the number of goals conceded is below the league average (62.3).

D. Performance Across Various Opponents

Assessing defender performance against different attacker types is also valuable for pre-game preparation. We categorise attackers by their teams' final 2023/24 EPL positions: top teams with higher quality players typically dominate possession, while lower-ranked teams typically play deeper and counterattack. Table IV compares the DP of top-rated center backs against each opposition type.

Most defenders record their highest DP against the top 1-5 teams, as DP tends to be higher when the ball is in high-threat areas. However, examples such as Max Kilman had the highest off-ball performance against lower-ranked teams. This analysis can help coaches assess player strengths and weaknesses. We extend this by evaluating Fabian Schär's influence on various center forwards by examining his mean DI when he has the highest attention score for an attacker out of all defenders. Table V lists the top and bottom three center forwards by mean DI against Schär, along with their total xT/90 (xT per 90 minutes), computed using the center forward's locations for all their on-ball attacking events in matches against Schär.

Interestingly, Fabian Schär recorded his lowest DI score against Erling Haaland, the league's top scorer during the 2023/24 EPL season. This analysis provides valuable insight into the specific attackers whom Schär demonstrated the greatest effectiveness in limiting ball reception.

TABLE IV
RANKING EPL DEFENDERS AND HOW WELL THEY PERFORM (DP) AGAINST EACH TYPE OF OPPOSITION. OPPOSITION TYPE IS SEPERATED BASED ON THE FINAL TEAM LEAGUE RANKING.

Defender	DP	1-5	6-10	11-15	16-20
F. Schär	0.0397	0.0436	0.0304	**0.0444**	0.0347
T. Gomes	0.0395	0.0433	0.0309	0.0348	**0.0490**
J. Andersen	0.0394	**0.0445**	0.0350	0.0359	0.0415
M. Kilman	0.0393	0.0398	0.0356	0.0389	**0.0431**
J. Tarkowski	0.0387	**0.0439**	0.0358	0.0380	0.0364
I. Zabarnyi	0.0386	**0.0431**	0.0418	0.0341	0.0319
J. Branthwaite	0.0384	**0.0441**	0.0374	0.0362	0.0336

E. Exploring Defender Performance

In this section, we analyse the DP metric of players under varying contexts. Firstly, we compare the DP of players when the ball is in different locations on the pitch in Figure 7.

The results show that the average DP is higher in dangerous pitch areas, as attackers pose greater threats, giving defenders with the same influence a higher DP. Thus, teams that defend frequently in high-risk areas tend to have higher DP scores. However, the link between DI and defensive actions (Section VI-F) indicates that DP increases in situations where the defending team is more likely to win the ball, giving insight into team defence beyond just the attacking threat. Interestingly, defender performance is notably better on the left wing compared to the right. This bias is similar to findings in [9], where players had higher on-ball decision ratings on the right wing for 2021/22 EPL data. Further analysis could explore if this trend holds for other seasons or leagues.

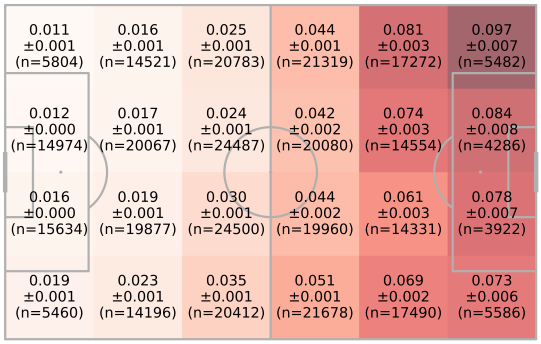

Fig. 7. Comparison of the average DP for each zone the ball occupies on the pitch. Teams attack from left to right.

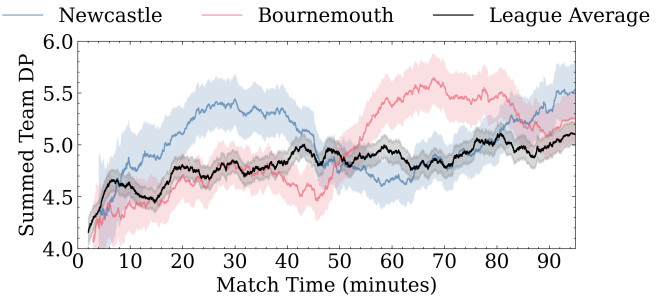

Fig. 8. Comparison of the combined team DP (rolling average) for varying match times for all games in our dataset, with shaded 95% confidence boundaries.

When comparing the average DP across different team roles, goalkeepers had the lowest mean DP (0.008), while wide defenders (0.053) and center midfielders (0.045) had the highest. This aligns with trends for tackles in the EPL, where these positions typically make the most tackles. We also analyse the combined DP for teams over different match times for all games in our dataset, comparing the league average with Newcastle and Bournemouth in Figure 8.

These results show that Newcastle's DP is above the league average in the first half, while Bournemouth's is higher in the second. T-tests found a statistically significant difference between the rolling averages of Bournemouth and Newcastle ($p<0.01$), indicating that they vary in their defensive patterns over time. This analysis helps clubs identify periods of weaker performance or increased defending for teams or players.

VIII. Discussion

Our framework introduces novel metrics to evaluate off-ball defensive contributions in football. While most research focuses on events, top teams use their positions to prevent events from happening, which is harder to measure. Merhej et al. [3] were the first to analyse on-ball defensive actions by predicting what was prevented. Our approach quantifies individual off-ball defensive performance for the first time, which players spend the vast majority of the match. Future work could extend the model to assess off-ball attacking contributions, such as runs or creating space, by analysing attacker attention and its influence on pass recipient probabilities. While GAPP uses velocity and acceleration features to represent player intentions and direction of play, future work could explore whether using sequences of past events as input enhances the model's understanding of temporal context further.

The GAPP model could also cluster similar play situations for players. Recent research compares past team situations [19], focusing on entire team structures. Using GAPP's node representations from the GAT layer outputs, we could generate latent representations of a player and their surroundings, enabling clustering of similar scenarios a player has faced, improving the efficiency of tactical analysis for clubs.

In future work, we plan to investigate alternative graph construction methods beyond the fully connected approach, such as proximity-based graphs, and conduct a computational analysis to compare the efficiency and effectiveness of these different graph structures for model performance and understanding true influence between players. Additionally, we plan to perform ablation studies to systematically evaluate the impact of various input features, hyperparameters, and architectural choices on model performance.

Football is a suitable testbed for our model as it is a data-rich, real-world environment, where tracking data provides many data points across games. In future work, we plan to test our approach across more leagues to capture a broader football context, as well as in other real-world multi-agent systems such as security scenarios or other team sports. For instance, in security games, the spatial positioning of defenders can influence attack locations, drawing similarities to the dynamics observed in football. Adapting the model to new domains will require thorough validation and potential adjustments, such as modifying the number of attention heads, layers or input data to reflect differences in relationships between agents. Additionally, the target variable in the GAPP model may need to be tailored to each domain. For example, by predicting the probability of an attack on each target in security settings.

While GAPP achieves state-of-the-art performance in football pass receiver prediction, its key value lies in interpreting off-ball defensive contributions. This interpretability supports post-match analysis, highlights team weaknesses, and introduces new metrics for player evaluation and recruitment. The model's explainable insights could make a substantial real-world impact by helping bridge the gap between domain expert football coaches and advanced machine learning, fostering trust and usability in data-driven decision-making.

IX. Conclusion

This paper introduces *GAPP*, a novel Graph Attention Network-based model for predicting pass reception probabilities in football. The model achieves a $\sim$6.4% $\pm$ 1.5% reduction in BCE loss compared to the best performing baseline while offering insights into off-ball defensive contributions, a critical yet underexplored area in football analytics. Derived from the GAPP attention mechanism, we introduce novel DI and DP metrics, providing new ways to evaluate off-ball defensive performance. A case study on the EPL highlights GAPP's ability to assess players and defender-attacker dynamics. We also show how explainable visualisations can improve trust in data-driven decision-making in football. Future work will explore the application of GAPP to other dynamic multi-agent systems, such as security and team sports.

Acknowledgment

We thank Gradient Sports for supporting and providing the data resources for this work. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. Gregory Everett was supported by Sentient Sports and Sarvapali Ramchurn was supported by the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1) and Responsible AI UK (EP/Y009800/1).

REFERENCES

[1] E. Shieh, B. An, R. Yang, M. Tambe, C. Baldwin, J. DiRenzo, B. Maule, and G. Meyer, "Protect: An application of computational game theory for the security of the ports of the united states," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012, pp. 2173–2179.

[2] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 1851–1861.

[3] C. Merhej, R. J. Beal, T. Matthews, and S. Ramchurn, "What happened next? using deep learning to value defensive actions in football event-data," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 3394–3403.

[4] J. Fernández, L. Bornn, and D. Cervone, "A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions," *Machine Learning*, vol. 110, no. 6, pp. 1389–1427, 2021.

[5] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[6] M. Cavus and P. Biecek, "Explainable expected goal models for performance analysis in football analytics," in *2022 ieee 9th international conference on data Science and advanced analytics (DSAA)*. IEEE, 2022, pp. 1–9.

[7] K. Singh, "Introducing expected threat (xt)," https://karun.in/blog/expected-threat.html, 2018, accessed: 2025-01-24.

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[9] P. Robberechts, M. Van Roy, and J. Davis, "un-xpass: Measuring soccer player's creativity," in *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, 2023, pp. 4768–4777.

[10] J. Fernández and L. Bornn, "Soccermap: A deep learning architecture for visually-interpretable analysis in soccer," in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Proceedings, Part V*. Springer, 2021, pp. 491–506.

[11] P. Rahimian, A. Oroojlooy, and L. Toka, "Towards optimized actions in critical situations of soccer games with deep reinforcement learning," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2021, pp. 1–12.

[12] G. Anzer, P. Bauer, U. Brefeld, and D. Faßmeyer, "Detection of tactical patterns using semi-supervised graph neural networks," in *16th MIT Sloan Sports Analytics Conference*, 2022.

[13] G. Everett, R. J. Beal, T. Matthews, T. J. Norman, and S. D. Ramchurn, "Optimising spatial teamwork under uncertainty," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2025, pp. 23 168–23 176.

[14] G. Everett, R. J. Beal, T. Matthews, J. Early, T. J. Norman, and S. D. Ramchurn, "Inferring player location in sports matches: Multi-agent spatial imputation from limited observations," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023, pp. 1643–1651.

[15] S. Omidshafiei, D. Hennes, M. Garnelo, Z. Wang, A. Recasens, E. Tarassov, Y. Yang, R. Elie, J. T. Connor, P. Muller *et al.*, "Multiagent off-screen behavior prediction in football," *Scientific reports*, vol. 12, no. 1, p. 8638, 2022.

[16] R. A. Yeh, A. G. Schwing, J. Huang, and K. Murphy, "Diverse generation for multi-agent sports games," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4610–4619.

[17] P. Rahimian, B. M. Mihalyi, and L. Toka, "In-game soccer outcome prediction with offline reinforcement learning," *Machine Learning*, vol. 113, no. 10, pp. 7393–7419, 2024.

[18] Z. Wang, P. Veličković, D. Hennes, N. Tomašev, L. Prince, M. Kaisers, Y. Bachrach, R. Elie, L. K. Wenliang, F. Piccinini *et al.*, "Tacticai: an ai assistant for football tactics," *Nature communications*, vol. 15, no. 1, p. 1906, 2024.

[19] M. Stöckl, T. Seidl, D. Marley, and P. Power, "Making offensive play predictable-using a graph convolutional network to understand defensive performance in soccer," in *Proceedings of the 15th MIT sloan sports analytics conference*, vol. 2022, 2021.

[20] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[21] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 10, pp. 4291–4308, 2020.

[22] A. BenTaieb and G. Hamarneh, "Predicting cancer with a recurrent visual attention model for histopathology images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 129–137.

[23] J. Simeunović, B. Schubnel, P.-J. Alet, R. E. Carrillo, and P. Frossard, "Interpretable temporal-spatial graph attention network for multi-site pv power forecasting," *Applied Energy*, vol. 327, p. 120127, 2022.

[24] X. Xing, F. Yang, H. Li, J. Zhang, Y. Zhao, M. Gao, J. Huang, and J. Yao, "An interpretable multi-level enhanced graph attention network for disease diagnosis with gene expression data," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 556–561.

[25] X. Li, L. Sun, M. Ling, and Y. Peng, "A survey of graph neural network based recommendation in social networks," *Neurocomputing*, vol. 549, p. 126441, 2023.

[26] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[27] J. Hu, L. Cao, T. Li, S. Dong, and P. Li, "Gat-li: a graph attention network based learning and interpreting method for functional brain network classification," *BMC bioinformatics*, vol. 22, pp. 1–20, 2021.

[28] J. Zhang, H. Li, P. Cheng, and J. Yan, "Interpretable wind power short-term power prediction model using deep graph attention network," *Energies*, vol. 17, no. 2, p. 384, 2024.

[29] Y. Zhang, Z. Zhou, Q. Yao, X. Chu, and B. Han, "Adaprop: Learning adaptive propagation for graph neural network based knowledge graph reasoning," in *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, 2023, pp. 3446–3457.

[30] Z. Zhou, J. Yao, J. Liu, X. Guo, Q. Yao, L. He, L. Wang, B. Zheng, and B. Han, "Combating bilateral edge noise for robust link prediction," *Advances in Neural Information Processing Systems*, vol. 36, pp. 21 368–21 414, 2023.

[31] Z. Zhou, Y. Zhang, J. Yao, Q. Yao, and B. Han, "Less is more: One-shot subgraph reasoning on large-scale knowledge graphs," *The Twelfth International Conference on Learning Representations.*, 2024.

[32] B. Bai, J. Liang, G. Zhang, H. Li, K. Bai, and F. Wang, "Why attentions may not be interpretable?" in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 25–34.

[33] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[34] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] W. Spearman, "Beyond expected goals," in *Proceedings of the 12th MIT sloan sports analytics conference*, 2018, pp. 1–17.

[36] Y. Dauxais and C. Gautrais, "Predicting pass receiver in football using distance based features," in *Machine Learning and Data Mining for Sports Analytics: 5th International Workshop, MLSA 2018, Co-located with ECML/PKDD 2018, Proceedings 5*. Springer, 2019, pp. 145–151.

[37] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[38] T. Zhao, D. Luo, X. Zhang, and S. Wang, "Towards faithful and consistent explanations for graph neural networks," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 634–642.

[39] Y.-M. Shin, S. Li, X. Cao, and W.-Y. Shin, "Faithful and accurate self-attention attribution for message passing neural networks via the computation tree viewpoint," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[40] Y. Liu, H. Li, Y. Guo, C. Kong, J. Li, and S. Wang, "Rethinking attention-model explainability through faithfulness violation test," in *International Conference on Machine Learning*. PMLR, 2022, pp. 13 807–13 824.

[41] R. Beal, N. Changder, T. Norman, and S. Ramchurn, "Learning the value of teamwork to form efficient teams," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7063–7070.