Analyzing Player and Team Chemistry in Soccer

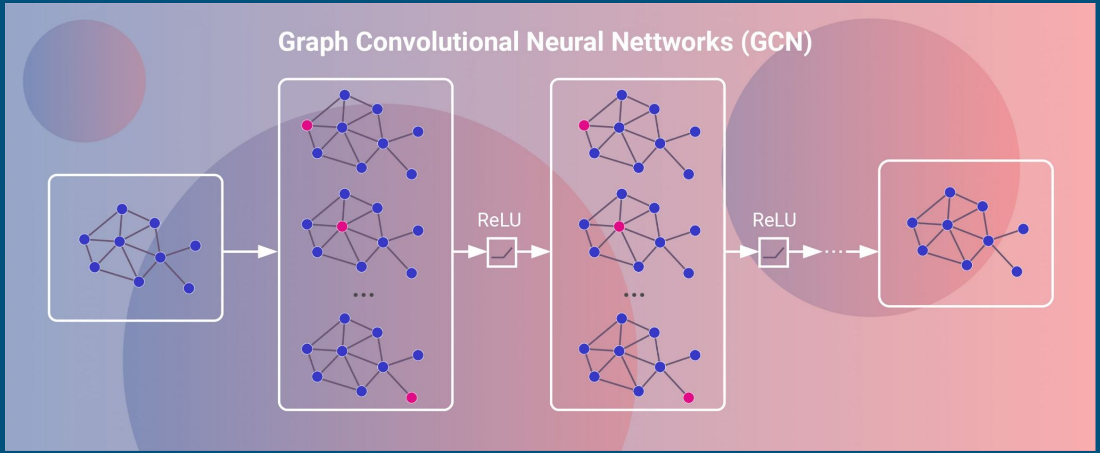
Tusheet Goli, Tejas Pradeep, Aman Jain, Akshay Pramod, Jeffrey Chang

Introduction

- Create a soccer team chemistry and performance evaluator
- Current approach is heuristic based (nationality and league), not data based
- Build a framework that can assess a team's projected performance based on various player and team attributes
  - Player attributes: pass completion percentage, goals per game, mentality
  - Team attributes: coach, formation, playing style, history

Applications

- Not many team chemistry based, data driven approaches
- New data and new technology (like GNNs) allows us to find additional insight
- Useful for soccer team management, gaming companies like FIFA and PES
- Enable dynamic team selection and scouting beyond individual statistics

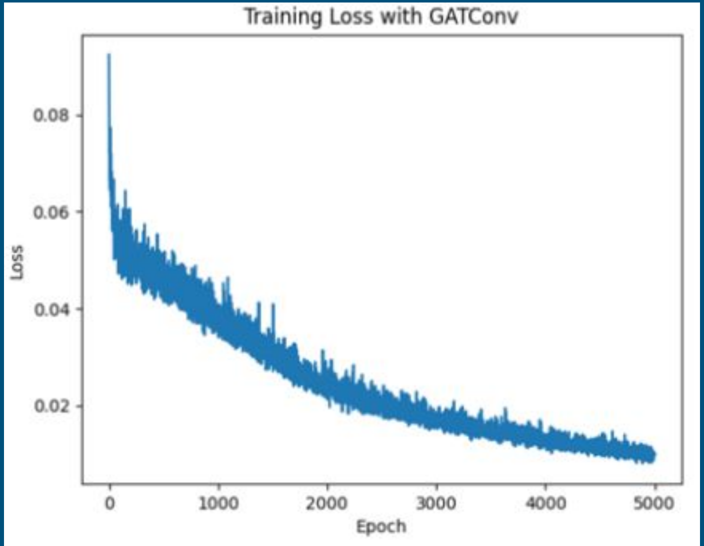


Previous Models

- We had three models from our previous implementation
  - GraphConv
  - GATConv
  - GCNConv
- Improved these models
  - Fine tuned them
  - Trained on newer and up-to-date datasets
- They serve as the baseline models

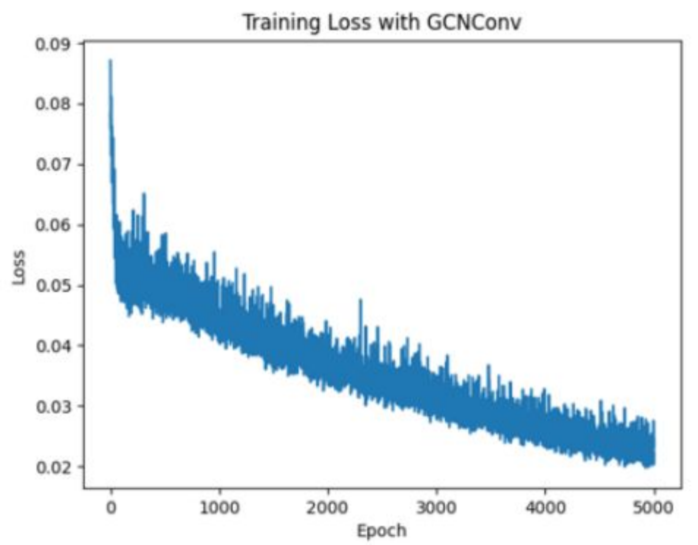
GATConv Model

- GATConv
  - GATConv incorporates a graph attention operator by utilizing the abilities of a attention network
  - Data Features Trained = ["position", "foot", "height_in_cm", "market_value_in_gbp"]
- Test MSE Error: 0.0185



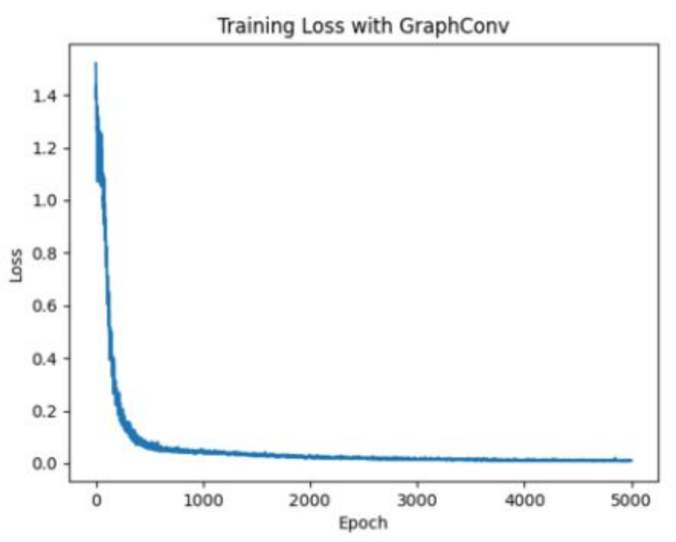
GCNConv Model

- GCNConv
    - GCNConv takes the weighted average of all it neighbouring nodes’ features and nodes with lower degrees get larger weights
    - Data Features Trained = ["position", "foot", "height_in_cm", "market_value_in_gbp"]
- Test MSE Error: 0.0136



GraphConv Model

- GraphConv
    - GraphConv is simple convolutional graphical network where you apply convolutions over a graph
    - Data Features Trained = ["position", "foot", "height_in_cm", "market_value_in_gbp"]
- Test MSE Error: 0.0251
- Best results of the 3 older models
- Used as baseline model



New Models

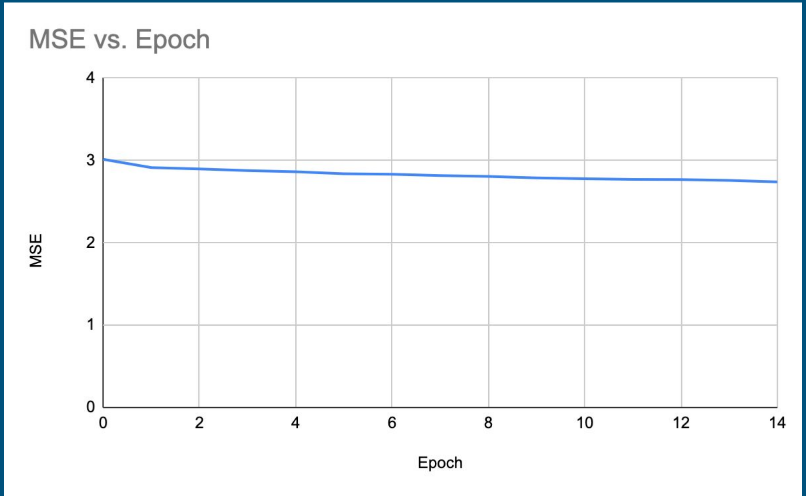
- We architected 3 new models to improve performance
  - Stat-Base Neural Network
  - TAGConv
  - ChebyShev
- Validated the model on our baseline
  - GraphConv

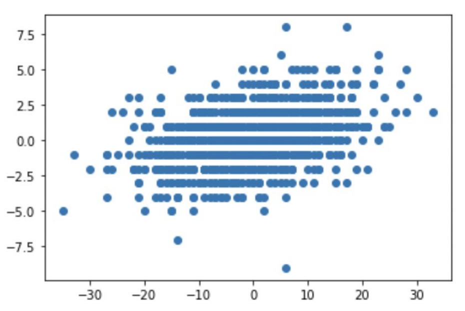
Model 1 - Stat-Based Neural Network

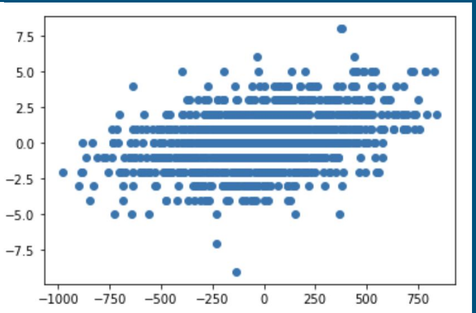
- **Overview:**
  - English Premier League ONLY (because of limited data set)
  - Use 40+ player & team statistics from games (i.e. assists, passes in the attacking third, tackles won, dribbles, etc.) to train a neural network
  - Network is trained on a dataset of ~1400 Premier league games over the last 6 years
    - Network is trained to predict goal difference between two teams
  - Inference is done on the aggregate statistics of the 11 players inputted
    - 11 player team provided is compared to the “statistically average” EPL team
    - Goal difference between team provided and Average Team is then normalized to a score between 0-100

Model 1 - MSE Convergence

- Converges rapidly, since the model is simple (3 layer linear neural network)
- MSE of ~ 2.8
    - This is in goal difference error - seems high because in practice, the model rarely predicts Goal Differences outside of the range [-1, 1], so outliers (which are not infrequent) increase MSE by a lot
    - Correlations are weak (graphs on right)
        - Shots on target (left)
        - Attacking third touches (right)







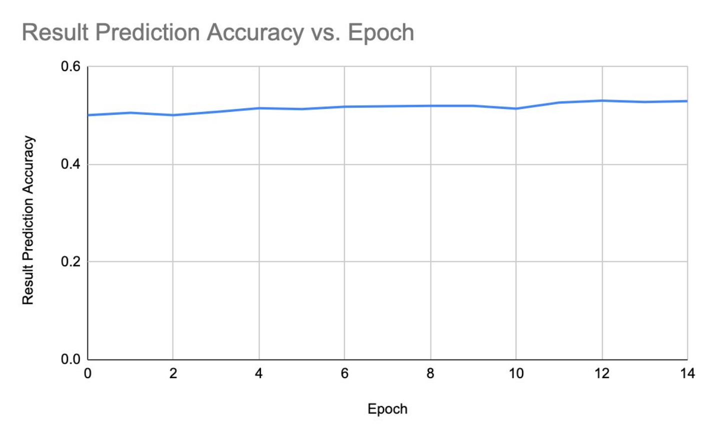
Model 1 - Simple Cases

- Teams that already exist:
  - Predicts 2017 Manchester City is a really good team (1st in EPL by 19 points)
    - Score: 83.8/100
  - Predicts 2020 Arsenal is slightly above average (8th in EPL)
    - Score: 55.6/100
  - Predicts 2021 Norwich City is a bad team (Last in EPL)
    - Score: 44.8/100

Model 1 - Challenging Cases

- Players out of position
  - Manchester City
    - Subtract two attackers, add two defenders
    - In reality should be imbalanced, but score is still high: 77.0/100
- No Goalie
  - Manchester City
    - Score nearly identical: 88.5/100
- Entirely one position group
  - Manchester City
    - All attackers (even without a goalie or defenders)
    - Score: 76.8/100
    - Linear model strongly correlates some variables a lot higher than other, especially goals scored and shots on target

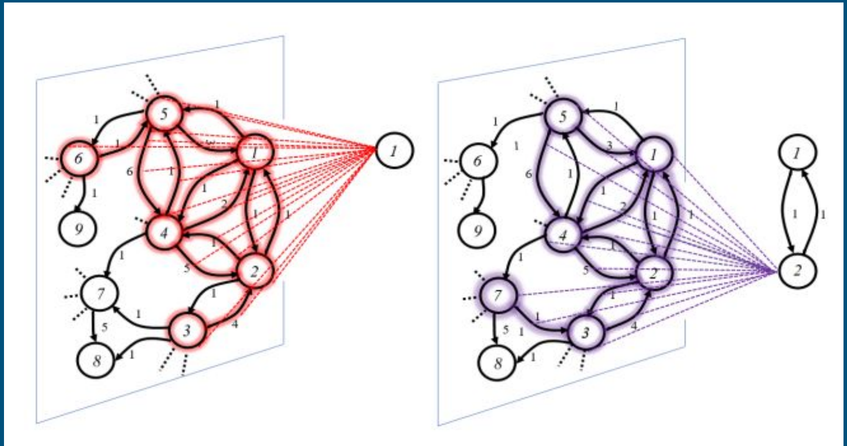
Model 1 - Best Use Cases



- Realistic team compositions
  - Useful for evaluating real trades that teams make
- Ran same model architecture with different evaluation criteria - **predicting game result** (Home Win, Home Loss, Draw)
  - Able to accurately predict 56.7% of games
  - The average Vegas predictions for 2021-2022 season were 57.9%
    - Average of 13 different oddsmakers

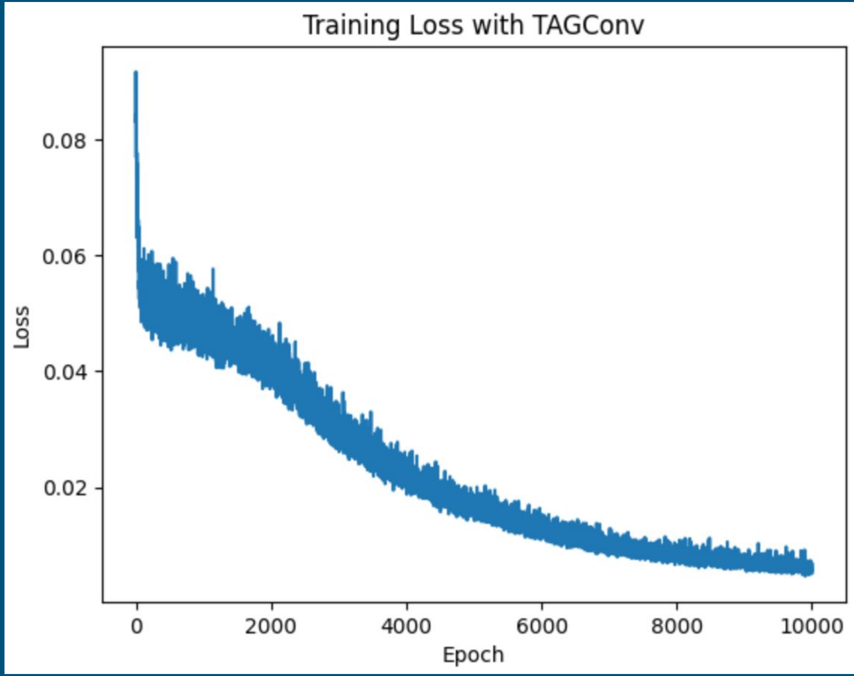
Model 2 - TAGConv

- The second model tested is the Topology Adaptive Graph Convolutional Network (TAGCN)
- Improvements on existing convolutional neural networks, and computationally simpler
- Two main types - spectral domain and vertex domain
- TAGCN creates polynomials of maximum degree 2 for each vertex using a similar convolution to traditional CNNs
- These polynomials can be accurately found using eigenvalue projections to avoid the loss of accuracy that normally occurs



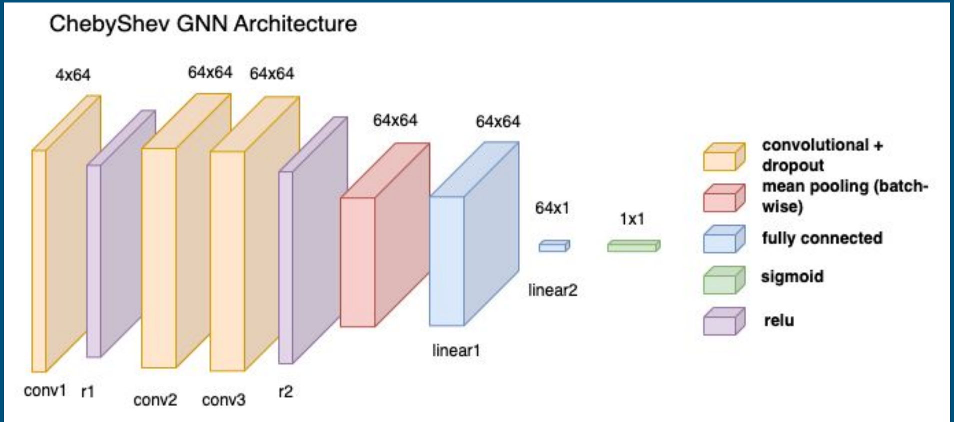
Model 2 - TAGConv MSE

- TAGConv Model Details
  - 5 TAGConv layers with ReLU activation functions
  - Adam optimizer with learning rate 1e-5
  - Mean Squared Error Loss
  - 10000 epochs
- Training Loss: 0.0041
- Test Loss: 0.0131



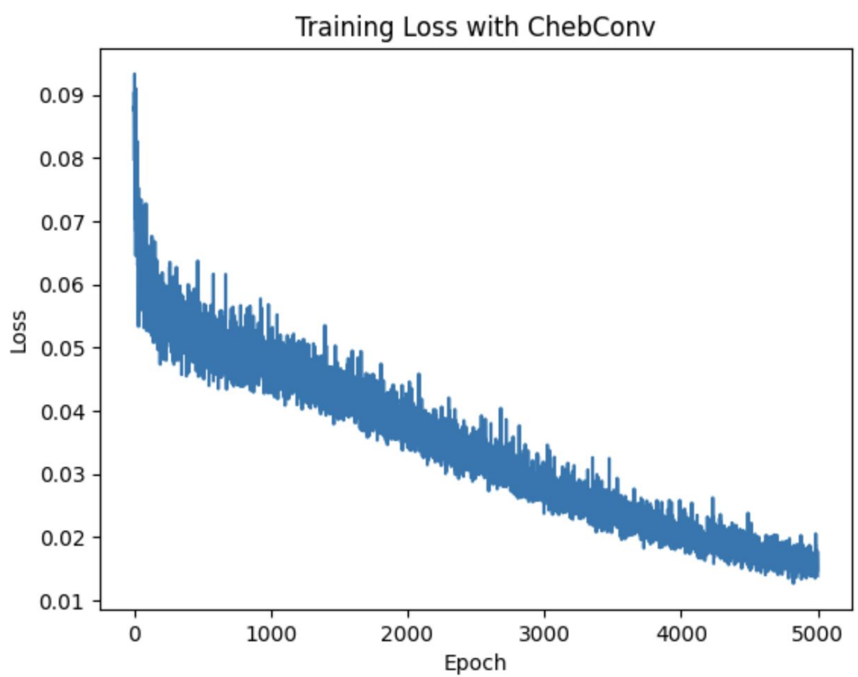
Model 3 - ChebyShev

- ChebyShev spectral graph convolution operator from “Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering”
    - Torch Geometric (GNN library)
- Model Features:
    - ChebyShev Filter Size (k) = 3
    - Data Features Trained = ["position", "height_in_cm", "market_value_in_gbp", “country”]



Model 3 - ChebyShev MSE

- Optimizer Parameters:
  - Adam Optimizer
  - Learning Rate = 1e-5
  - Loss Criterion: Mean Squared Error (MSE) Loss
- Test MSE Error: 0.014



Easy Classification Tasks

- Case: Running the model on an existing well performing team
  - Test for if the model is performing correctly
- Example:
  - Consider the current Champions League champions, Real Madrid
  - Any model used to predict team performance and chemistry must have a high performance prediction for Real Madrid
  - Similar approach can also be used for other clubs and national teams that have a good performance history

Easy Classification Tasks

- **Case: Run on Real Madrid base/starting team**
- Original Model
  - Strength Score: 84.162/100
- Model 2 - TAGConv
  - Strength Score: 74.617/100
- Model 3 - ChebyShev
  - Strength Score: 93.598/100

Medium Classification Tasks

- Lewandowski as goalkeeper
  - Lewandowski is Barcelona's current striker
  - Playing him out of position should lead to a worse team
  - Expected: Significantly worse results with the change
  - Tests the model's ability to perceive position
- Griezmann to PSG
  - Griezmann is a world class French player
  - Consider a transfer to french side PSG to replace Neymar, a better player
  - Expected: Team becomes slightly worse with the change
  - Tests the model's reliance on chemistry over performance
- Ronaldo to Manchester United
  - Ronaldo is a world class striker who has a history of playing with Manchester United
  - This recent transfer was expected to have good results
  - In reality the transfer lead to suboptimal results
  - Expected: Model predicts worse results for the team
  - Tests the model's ability to see hidden attributes that can predict a negative result

Medium Classification Tasks

- **Case 1: Put Lewandowski as goalkeeper for Barcelona**
- **Case 2: Add Griezmann to PSG (replace an objectively stronger player in Neymar)**
- **Case 3: Add Ronaldo to Manchester United as a striker**
- Original Model
  - (1) Before: 77.631/100, After: 86.974/100
  - (2) Before: 80.351/100, After: 80.511/100
  - (3) Before: 80.344/100, After: 65.976/100
- Model 2 - TAGConv
  - (1) Before: 80.301/100, After: 22.345/100
  - (2) Before: 78.943/100, After: 78.682/100
  - (3) Before: 78.943/100, After: 82.828/100
- Model 3 - ChebyShev
  - (1) Before: 88.696/100, After: 39.638/100
  - (2) Before: 94.73/100, After: 94.57/100
  - (3) Before: 94.73/100, After: 94.68/100

Hard Classification Tasks

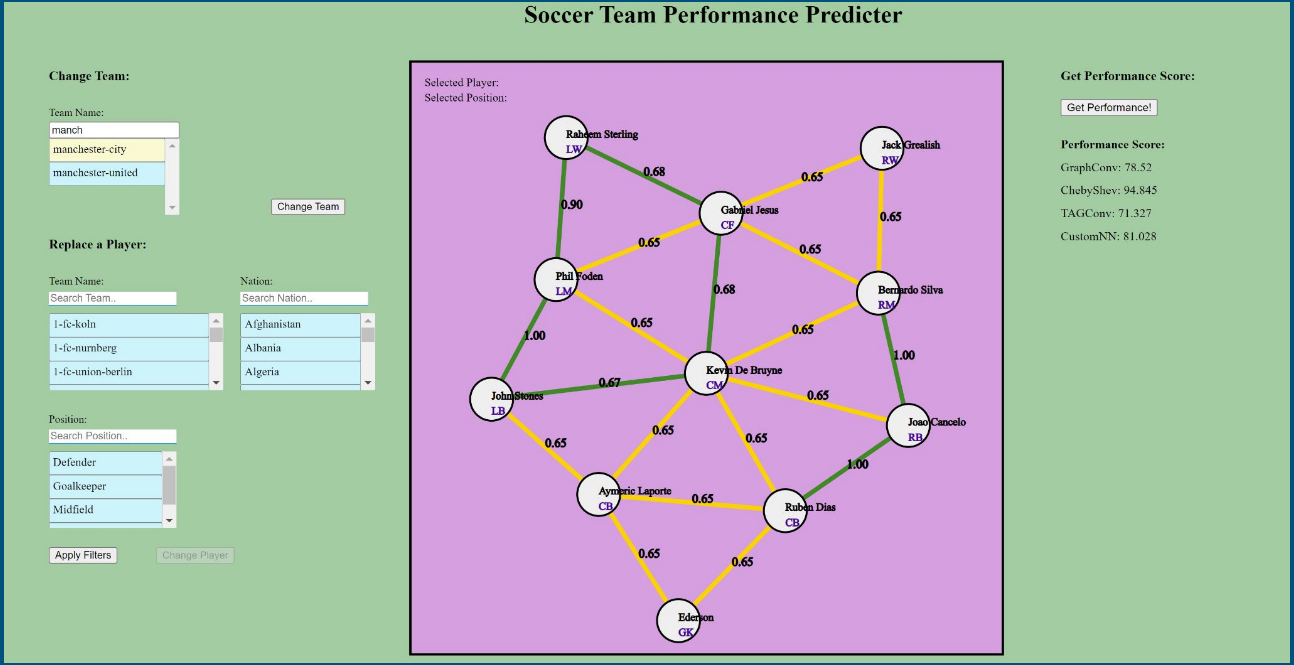
- Case: If there is little data and history between players when considering a new player for a team
  - A situation that is very difficult to predict due to the lack of data and uncertainty
- Example:
  - Additional of an established player to a good squad, but the player impacting the squad in an an unreliable way
  - Transfer of Lewandowski to Barcelona

Hard Classification Tasks

- **Case: Add Lewandowski to Barcelona as a striker**
- Original Model
  - Before: 77.656/100, After: 73.638/100
- Model 2 - TAGConv
  - Before: 80.301/100, After: 85.054/100
- Model 3 - ChebyShev
  - Before: 88.696/100, After: 90.098/100

Viz Tool

- Data:
  - MySQL (.sqlite)
- Backend:
  - Python Flask
- Frontend:
  - ReactJS
- Database Hosting:
  - Heroku DB



Thank you!