Tennis Match Prediction: A Comparative Study of GNN and Feature-Based Models

Donato Festa

OUTLINE

❑ The problem: Why is Tennis Prediction Challenging?

❑ Dataset & Preprocessing

❑ Proposed Approaches:

- ❖ XGBoost
- ❖ GNN

❑ Evaluation Protocol

❑ Results & Analysis

❑ Conclusions & Future Work

The Problem: Tennis Match Prediction

Complex Interplay: 🧩

✓ skill, physical form, psychology, surface.

Traditional Models' Limitations: 📦

✓ Rely on aggregated stats, treat matches as independent.

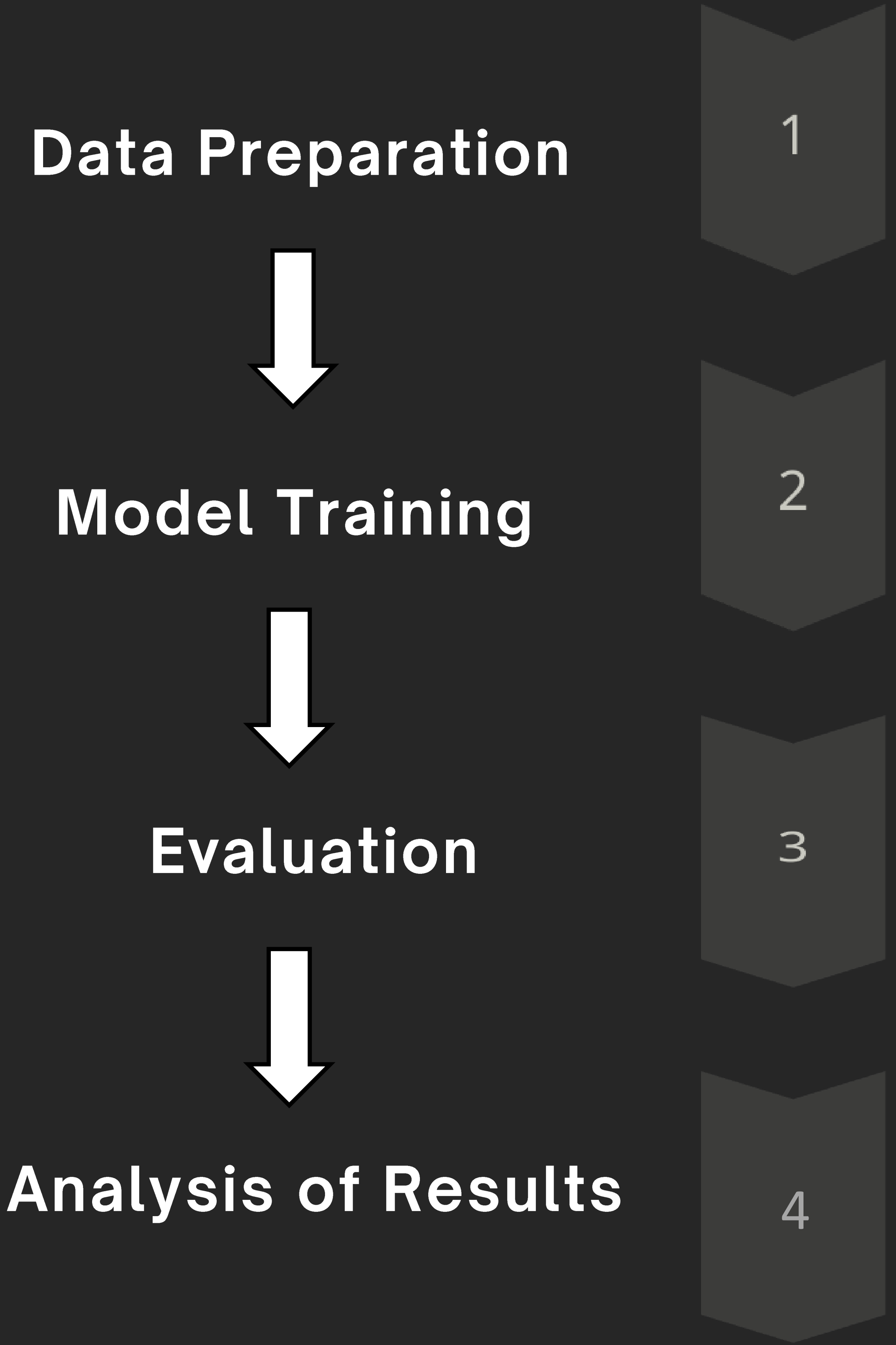
Fundamental Truth: 🕸️

✓ Player strength is **relational** (who they beat/lost to).

Project Goal: 🎯

✓ Compare feature-based vs. graph-based models to capture this relational structure.

Experimental Pipeline



1 Data Preparation

2 Model Training

3 Evaluation

4 Analysis of Results

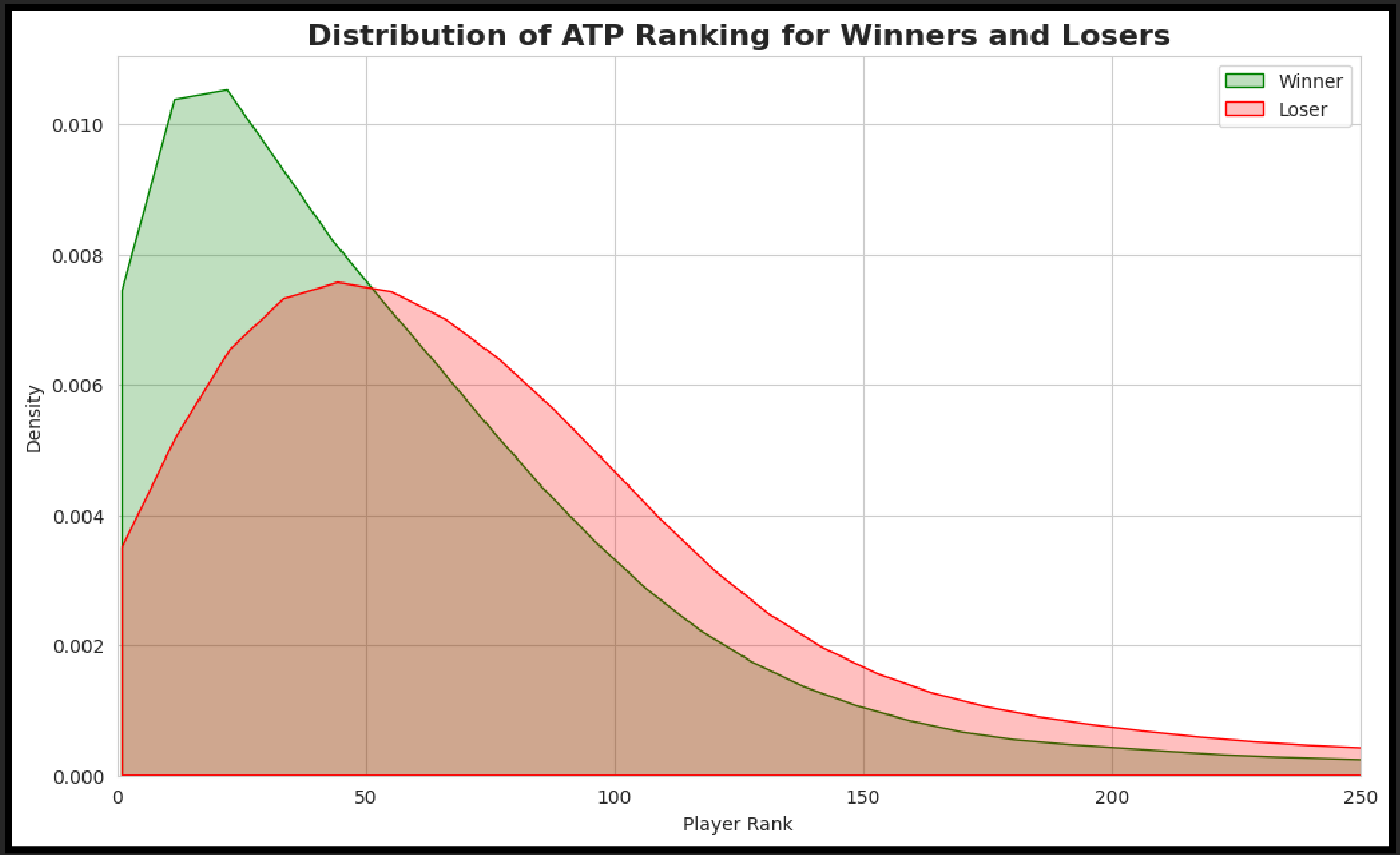
Dataset & Preprocessing

Source: Jeff Sackmann’s ATP matches (2000-2023).

- 2000 - 2023
- >70.000 matches

Key Preprocessing: Chronological sorting.

EDA Insight: Winner/Loser Rank Distribution



Distribution of ATP Ranking for Winners and Losers

Proposed Approaches

We explored two paradigms:

1

XGBoost Baseline (Feature-Based)

- Traditional ML classifier for tabular data.
- Relies on hand-crafted features.

2

GNN Model (Graph-Based)

- Deep Learning on Graph-structured data.
- Learns representations (embeddings) from graph structure.
- Frames problem as Link Prediction.

Baseline Model: XGBoost with Engineered Features

□ **Key Engineered Features (Differential):**

- Elo rating
- Rank
- Head-to-Head
- Winning Streak
- Age
- Fatigue
- Surface (one-hot encoded)

□ **Data Structuring:** Randomized Player1/Player2 assignment to prevent bias.

□ **Regularization:**

- max_depth
- subsample
- gamma
- colsample_bytree
- learning_rate

Graph-Based Model: GNN for Link Prediction

□ Graph Construction:

- Nodes = Players
- Directed Edges = Winner -> Loser

□ Model Architecture *(GNNLinkPredictor)*:

✓ **Encoder (GraphSAGE):** 2 layers, learns player embeddings from scratch.

✓ **Decoder:** Dot product (z_u * z_v) for link plausibility.

□ Traning:

✓ **Negative Sampling:** Generates false links to teach model what not to predict.

✓ **Minimizes** Binary Cross-Entropy Loss + L2 Regularization

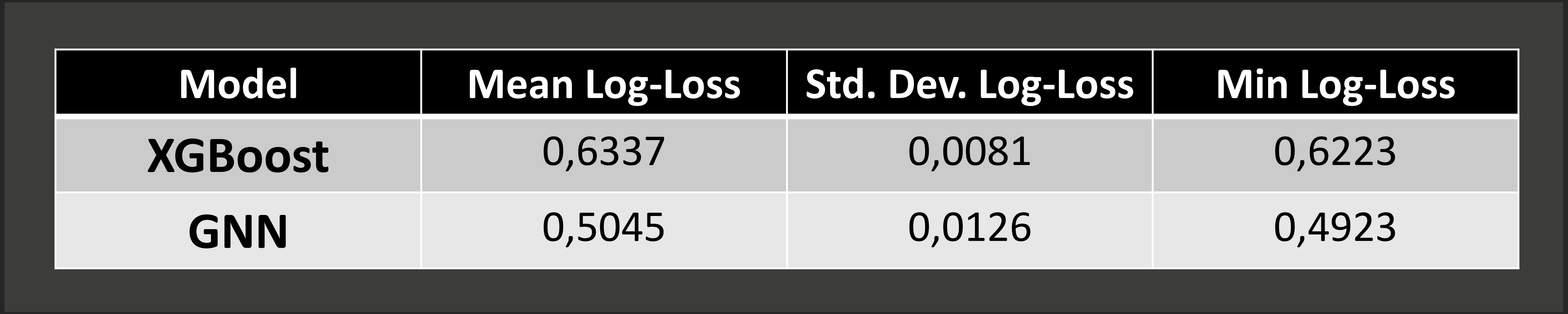
✓ **Full-Batch Traning:** Entire graph processed per epoch.

Evaluation Protocol

Identical & Rigorous for Both Models:

- **Temporal Cross-Validation (5 folds):**
  - ❖ Always train on past to predict future
- **Multiple Runs (10 times):**
  - ❖ Different hyperparameter configurations
    - ✓ XGBoost: Randomized Search
    - ✓ GNN: Predefined Grid
- **Primary Metric: Log-Loss**
- **Statistical Test:** t-test on Log-Loss scores

Quantitative Results



Model	Mean Log-Loss	Std. Dev. Log-Loss	Min Log-Loss
XGBoost	0,6337	0,0081	0,6223
GNN	0,5045	0,0126	0,4923

❑ **GNN Superiority:** Substantially lower Mean Log-Loss.

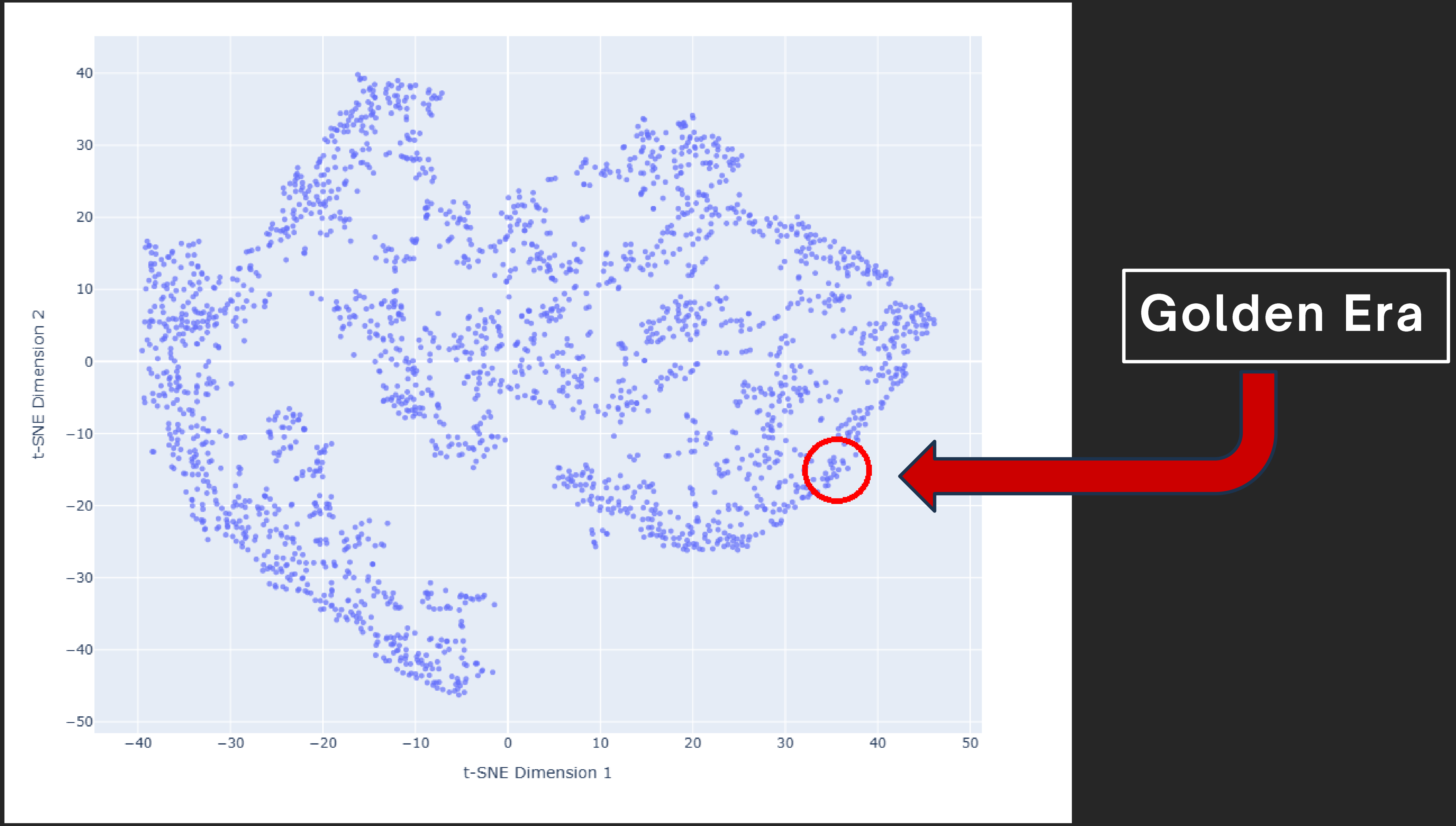
❑ **Statistical Significance:**

- ✓ t = 36,66
- ✓ p ≈ 4.1 x 10^{-11} ($p < 0.001$).

❑ **Conclusion:** GNN’s improvement is not due to random chance.

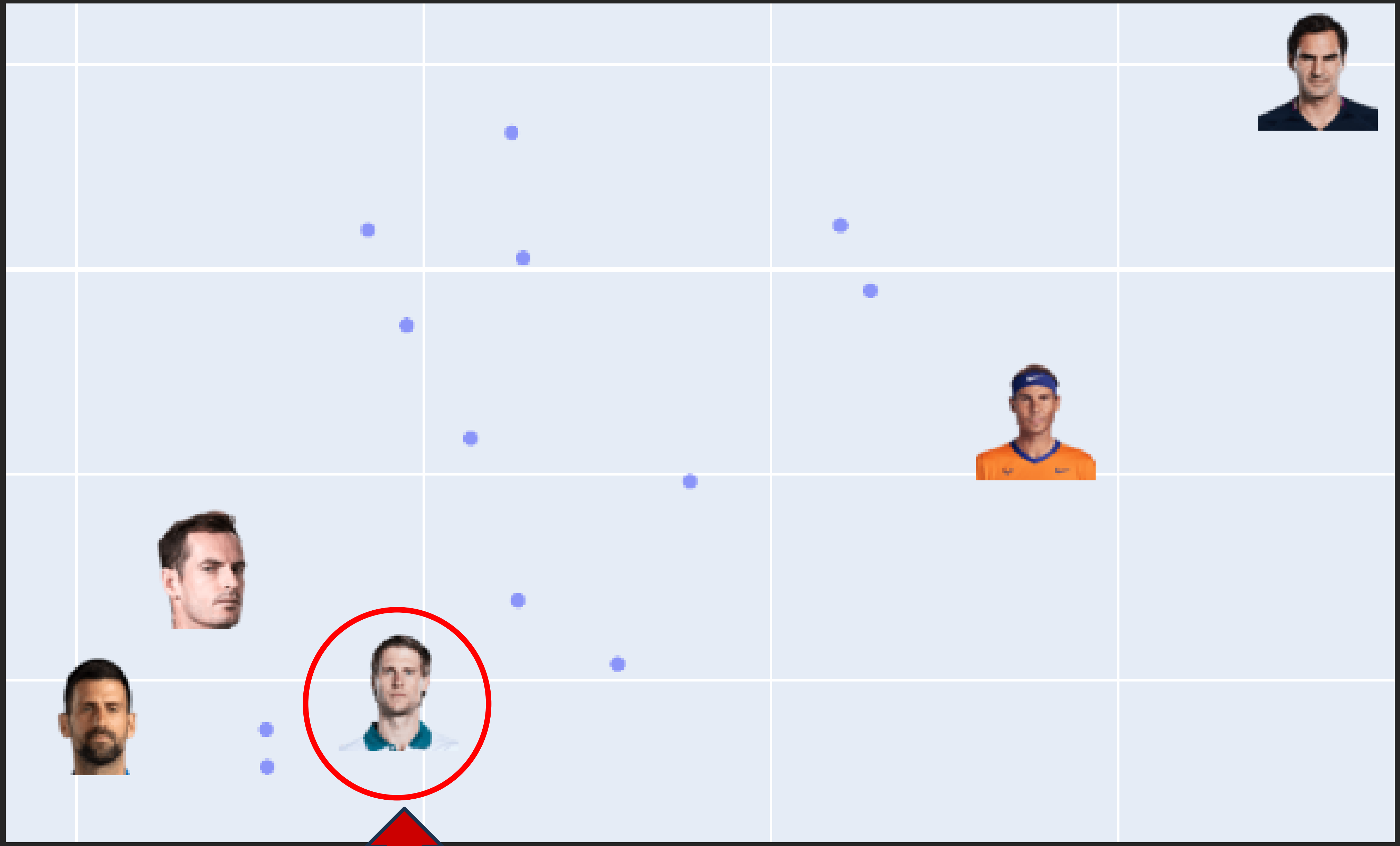
Qualitative Analysis: GNN Embeddings

t-SNE Visualization: Projects 64D embeddings to 2D.



Golden Era Analysis

The model also places other players from that era within this cluster.



Andreas Seppi

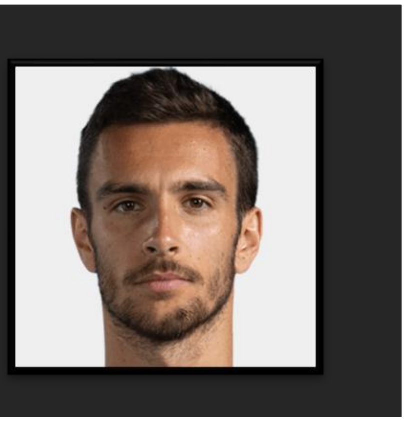
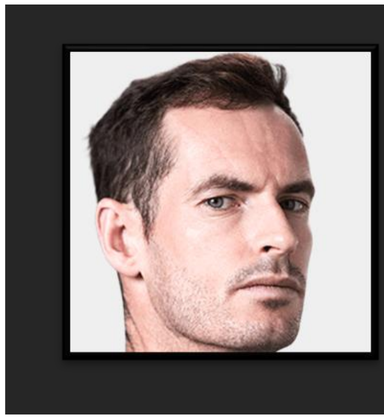
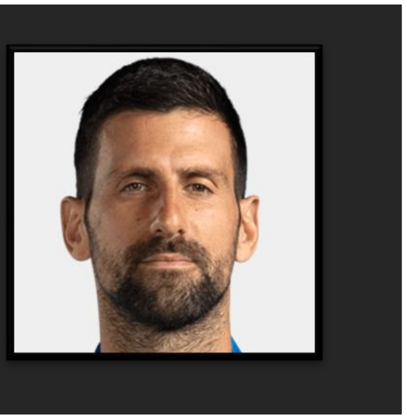
Opponent	Matches Played
Roger Federer	15
Novak Djokovic	12
Rafael Nadal	9
Andy Murray	9
Lleyton Hewitt	7
Gael Monfils	7
Feliciano Lopez	4
Andy Roddick	2

✓ Also has 10 top 10 wins.

Real-World Prediction Example (GNN)

Matches:

Winner:

	Felix Auger-Aliassime vs. Lorenzo Musetti			66,08%
	Carlos Alcaraz vs. Jannik Sinner			50,18%
	Andy Murray vs. Novak Djokovic			99,99%

Note: the model was trained with matches until 2022.

Conclusions & Key Takeaways

1. **GNN Superiority:** Statistically significant improvement over XGBoost.

2. **Relational Learning:** GNN captures structures like reputation and context.

3. **Powerful Paradigm:** Graph-based approach is highly effective.

4. **Identified Limitations:** Over-confidence in prediction for historical matchups.

Limitations & Future Work

LIMITATIONS

1. Over Confidence 🧐
2. Limited to ATP circuit ♂️

FUTURE WORK

1. Advanced Decoders 🧠
2. Richer Graph Features 🌳
3. Hybrid Models 🧪
4. Temporal Graph Networks (TGNs) 🔄

Thank You For Your Attention! 🚀

Donato Festa