Markov Chain Aggregation of Strategy Evolution in Football RL Agents

Roman Dupraz–Bardou
ENS Paris-Saclay, MVA Master
Gif-sur-Yvette, France
ENSAE Paris
Palaiseau, France
roman.dupraz-bardou@ensae.fr

Samuel Marthely
ENS Paris-Saclay, MVA Master
Gif-sur-Yvette, France

Abstract

We apply Markov Chain Aggregation (MCA) to the study of strategy evolution in a football reinforcement learning agent trained in the Google Research Football environment (GRF). Rather than tracking performance through reward curves alone, we represent each checkpoint by a behavioural feature vector covering reward, possession, spatial progression, action frequencies and pass network metrics, and model the training trajectory as a Markov chain over micro-states. Aggregating these into macro-states via a lumpability-validated partition reveals three persistent strategic regimes: sterile possession, unstable offensive transition, and direct attacking policy. A chi-squared lumpability test does not reject the aggregation ($X^2 = 1.86$, $p = 0.40$), supporting a Markovian interpretation of the coarse-grained dynamics. To our knowledge, this is the first application of MCA to strategy evolution in a football RL environment, and the results demonstrate that the approach can turn a sequence of training checkpoints into an interpretable dynamical model of emergent behaviour.

CCS Concepts: • **Computing methodologies** → **Markov decision processes**; *Reinforcement learning*; *Multi-agent systems*; • **Theory of computation** → Stochastic processes.

Keywords: Markov chain aggregation, agent-based modelling, reinforcement learning, strategy evolution, emergent behaviour, Google Research Football, *lumpability*

Roman Dupraz–Bardou and Samuel Marthely. 2026. Markov Chain Aggregation of Strategy Evolution in Football RL Agents . In *Proceedings of MVA's Interactions Course (INT2026MVA).* ACM, New York, NY, USA, 6 pages.

INT2026MVA, Gif-sur-Yvette, France

1 Introduction

Sports analytics has its roots in American sports. From baseball to basketball and American football, the philosophy is simple: everything that can be measured, should be. Over the last two decades, this data-driven culture has made its way into association football, mostly in the English Premier League. Liverpool FC is perhaps the most emblematic example, having built a sustained competitive advantage through analytics under the guidance of Ian Graham, who recently documented the club's philosophy in [3]. Yet real football data remains expensive, noisy and hard to collect at scale. This naturally motivates the use of virtual environments for simulation. It offers a controlled setting where strategies can be tested, statistics inferred and hypotheses validated at zero marginal cost. The idea is not new, football video games such as EA Sports FC (formerly FIFA), Konami eFootball (formerly PES) or Football Manager have long served as informal simulators, with the latter even finding its way into professional clubs and managers' workflows[1].

Meanwhile, deep reinforcement learning has demonstrated remarkable capabilities in complex sequential decision-making tasks from AlphaGo mastering the game of Go [13] to OpenAI Five defeating professional Dota 2 players [9]. The Google Research Football environment [8] brings this trend to football analytics, providing an open-source research sandbox built on top of the Gameplay Football engine. It allows agents to be trained on a range of scenarios from simple academy drills to full 11v11 matches and opens new possibilities for studying emergent tactical behaviour in a reproducible setting. Scott et al. [12] took a first step in this direction, showing through social network analysis that sufficiently competitive RL agents spontaneously develop strategies similar to those of real football players. Yet their analysis remains largely descriptive: strategy evolution is tracked by plotting continuous metrics against a competitiveness signal, without asking whether this evolution is smooth and gradual or whether it proceeds through discrete regimes separated by sharp transitions.

In this paper, we address this gap and ask: *does strategy evolution during training follow smooth, continuous dynamics, or does it proceed through discrete behavioural regimes ?*

[1] https://en.wikipedia.org/wiki/Football_Manager#Influence

Figure 1. A full 11v11 match in the Google Research Football environment [8].

To answer it, we apply Markov Chain Aggregation (MCA) [1] to the training trajectory of a GRF agent, turning a qualitative narrative into a rigorous dynamical model. To our knowledge, this is the first work to apply Markov Chain Aggregation to model strategy evolution in a football reinforcement learning environment.

2 Related Work

This work builds on the Google Research Football environment (GRF), introduced by Kurach et al. [8] as an open-source reinforcement learning benchmark based on the open-source game Gameplay Football. As illustrated in Figure 1, the environment provides a physics-based 3D football simulation with a real-time graphical rendering, designed to address key limitations of existing RL environments such as being too easy to solve, computationally prohibitive or lacking stochasticity. GRF offers a natural balance between short-term control, where agents learn concepts such as passing and strategy. It provides three main components: the Football Engine, a highly-optimised game engine, the Football Benchmarks, a set of tasks of varying difficulty and the Football Academy, a collection of progressively harder scenarios ranging from simple one-player drills to full 11v11 matches. State-of-the-art Deep-RL algorithms including PPO [11], IMPALA [2] and Ape-X DQN [5] are evaluated as baselines, establishing GRF as a demanding benchmark for studying emergent behaviour in competitive settings.

Scott et al. [12] provide the closest prior work to our own. Their framework trains PPO agents [11] at increasing competitiveness levels against built-in bots, ranks them via the TrueSkill rating system [4] (a Bayesian skill rating framework) then extracts pass and shot event data from simulated matches to apply Social Network Analysis (SNA) on the resulting pass networks. They computed classic graph metrics such as closeness centrality, betweenness centrality and PageRank on the weighted directed pass network. Their main finding is that sufficiently competitive agents spontaneously develop a well-balanced passing strategy similar to that of real-world football teams, without being explicitly programmed to do so. They obtain a near-perfect negative correlation for minimum PageRank ($r = -0.91, p = 0.001$) and a strong positive correlation for betweenness standard deviation ($r = 0.65, p = 0.009$).

However, strategy evolution is only tracked by plotting continuous SNA metrics against a competitiveness signal, without asking whether this evolution is smooth and gradual or whether it proceeds through discrete behavioural regimes making it mainly a descriptive analysis. The mechanisms underlying the observed emergent alignment are left unmodelled and the authors themselves acknowledge difficulties in applying state-of-the-art football analysis methods due to mismatches between the simulation's data representation and real-world formats. It is precisely this gap, between descriptive correlation and dynamical modelling, that our work addresses.

To bridge it, we turn to the Agent-Based Modelling (ABM) framework. ABMs are computational models that simulate the behaviour of individual agents and observe how macro-level patterns emerge from their local interactions [1]. Originally rooted in cellular automata, from von Neumann's self-reproducing machines to Conway's Game of Life, ABMs have since become a standard tool across the social sciences, economics, epidemiology and biology to study emergent collective behaviour. In our setting, a football RL agent undergoing training can naturally be cast as such a system, where its behavioural state evolves over time and the sequence of states across training checkpoints defines a trajectory amenable to formal dynamical analysis. Markov Chain Aggregation (MCA), as formalised by Banisch [1], provides a principled framework for coarse-graining such a high-dimensional behavioural trajectory into a smaller number of interpretable macro-states and the transition probabilities between them.

To check that an aggregation preserves the Markov property, Jernigan and Baran [6] propose a chi-squared test of *lumpability*. A finite Markov chain $\{X_t\}$ on state space $S = \{1, \dots, n\}$ is said to be *strongly lumpable* with respect to a partition $S = \{L^{(1)}, \dots, L^{(m)}\}$ if the lumped chain $\{Y_t\}$, obtained by replacing each state with the index of its lump, is itself a Markov chain. The necessary and sufficient condition, based on Kemeny and Snell [7], is that the state-to-lump transition probabilities

$$v_{kj} = \sum_{l \in L^{(j)}} p_{kl}, \quad k \in S,\ j = 1, \dots, m \tag{1}$$

must take the same value for every $k \in L^{(i)}$. When this holds, there exists an $m \times m$ lump-to-lump transition matrix $Q = (q_{ij})_{i,j}$ such that $v_{kj} = q_{ij}$ for all $k \in L^{(i)}$.

To test this hypothesis from a single finite realisation, Jernigan and Baran [6] substitute maximum likelihood estimators for the transition probabilities and construct the chi-squared statistic

$$X^2 = \sum_{k=1}^{n} \sum_{j=1}^{m} \frac{\left(o_{kj} - n_{k\bullet} m_{ij}/m_{i\bullet}\right)^2}{n_{k\bullet} m_{ij}/m_{i\bullet}} \tag{2}$$

where o_{kj} is the observed number of transitions from state k to lump j, $n_{k\bullet}$ is the total number of transitions out of state k, and m_{ij} is the lump-to-lump transition count. Under the null hypothesis of *lumpability*, X^2 follows asymptotically a χ^2 distribution with $(m-1)(n-m)$ degrees of freedom. A partition is rejected at level α when the corresponding p-value falls below α.

3 Methodology & Experimental Setup

Our code and experiments are fully reproducible and available here. The pipeline proceeds in five steps, as described below.

(a) `academy_pass_and_shoot_with_keeper` (our scenario) **(b)** `academy_3_vs_1_with_keeper`

Figure 2. Examples of Football Academy scenarios from the GRF environment [8].

Training. A single PPO agent [11] is trained on the academy_pass_and_shoot_with_keeper scenario (Figure 2) using Stable-Baselines3 [10] with 8 parallel environments. A reward combining goal scoring and intermediate checkpoints, and the `simple115v2` state representation over 500k timesteps (all hyperparameters are in Table 1). Model checkpoints are saved every 10k steps, yielding 55 behavioural snapshots.

Feature extraction. Each checkpoint is evaluated over 50 deterministic episodes. Beyond reward statistics, we extract a behavioural feature vector inspired by Scott et al. [12], comprising performance metrics (mean reward, episode length, goal success rate), possession indicators (own possession share, possession losses and regains per episode), spatial progression features (ball position, forward progress, offensive third share), action frequencies over the 19 available actions, and pass network metrics (pass rate, pass network density, ball carrier entropy, unique ball carriers). Together, these features characterise the evolution of the team's playing strategy across training.

Micro-state discretisation. The feature vectors are standardised and clustered with k-means. The number of micro-states is chosen by inspecting the spectral gap of the normalised Laplacian of the RBF feature similarity matrix, which suggests 4 micro-states. However, it is really a descriptive heuristic because there is not a clear rule for choosing this gap.

Transition matrix estimation. An empirical $n \times n$ transition count matrix is estimated from the ordered sequence of micro-states across checkpoints, then row-normalised to obtain $\hat{P}_{\text{micro}}$ (the transition matrix between micro-states).

Macro-state aggregation. Candidate macro-state partitions are obtained by applying k-means on the rows of $\hat{P}_{\text{micro}}$ for $m \in \{2, 3\}$. Each partition is validated via the *lumpability* test of Jernigan and Baran [6], and the partition with $m = 3$ macro-states is retained ($p = 0.3951$).

4 Results

We now report the outcome of the pipeline described above. We first analyse the micro-state dynamics, then present the macro-state aggregation and its lumpability validation, before providing a behavioural interpretation of the identified regimes.

Micro-state dynamics. The checkpoint trajectory is first discretised into 4 micro-states. The empirical transition count matrix is

$$N_{\text{micro}} = \begin{pmatrix} 3 & 0 & 1 & 1 \\ 0 & 14 & 4 & 0 \\ 1 & 5 & 14 & 1 \\ 0 & 0 & 2 & 3 \end{pmatrix},$$

which gives the row-normalised transition matrix

$$P_{\text{micro}} = \begin{pmatrix} 0.600 & 0.000 & 0.200 & 0.200 \\ 0.000 & 0.778 & 0.222 & 0.000 \\ 0.048 & 0.238 & 0.667 & 0.048 \\ 0.000 & 0.000 & 0.400 & 0.600 \end{pmatrix}.$$

The diagonal terms are relatively high, especially for micro-states 1 and 2, with self-transition probabilities of 0.778 and 0.667. This indicates that the agent does not move randomly between behavioural profiles: several checkpoints remain in the same local regime before transitioning to another one. This supports the hypothesis that training induces persistent behavioural regimes rather than a purely smooth or noisy evolution.

Macro-state aggregation. The micro-states are then aggregated into 3 macro-states by clustering their transition profiles, i.e. the rows of P_{micro}. The obtained mapping is

$$0 \mapsto 2, \qquad 1 \mapsto 0, \qquad 2 \mapsto 1, \qquad 3 \mapsto 1.$$

The resulting macro-state transition count matrix is

$$N_{\text{macro}} = \begin{pmatrix} 14 & 4 & 0 \\ 5 & 20 & 1 \\ 0 & 2 & 3 \end{pmatrix},$$

and the associated transition matrix is

$$P_{\text{macro}} = \begin{pmatrix} 0.778 & 0.222 & 0.000 \\ 0.192 & 0.769 & 0.038 \\ 0.000 & 0.400 & 0.600 \end{pmatrix}.$$

This macro-level description reveals three persistent coarse regimes. The three macro-states differ strongly in reward, success rate, episode length, and action profile. Detailed averages are reported in Appendix B.

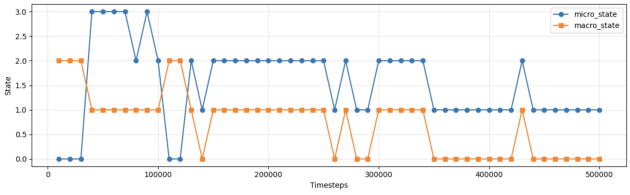

Figure 3. Evolution of micro-state and macro-state assignments along the training trajectory.

The temporal organisation of these regimes is shown in Figure 3. The early checkpoints are mostly assigned to macro-state 2, then the trajectory moves towards macro-state 1, which dominates the intermediate part of training. Macro-state 0 appears from approximately $140k$ steps and becomes increasingly frequent towards the end of training. The evolution is not perfectly monotonic, since the agent alternates between macro-states 1 and 0 during the middle of training but the overall pattern is structured: the training trajectory moves from a sterile-possession regime to an unstable offensive transition regime, and finally to a more efficient vertical attacking policy.

Lumpability check. To assess whether the aggregation is compatible with a Markovian coarse-graining, we apply the chi-squared *lumpability* test of Jernigan and Baran [6]. For the 3-macro-state partition, the test gives

$$X^2 = 1.8571, \qquad \text{df} = 2, \qquad p = 0.3951.$$

Since $p > 0.05$, the null hypothesis of *lumpability* is not rejected. This does not prove exact *lumpability* but it indicates that the proposed aggregation is statistically compatible with a Markovian macro-description on this finite trajectory. A 2-macro-state partition is also not rejected ($X^2 = 2.2831$, $\text{df} = 2$, $p = 0.3193$) but the 3-state partition gives a more interpretable tactical decomposition.

Behavioural interpretation of the macro-states. The three macro-states correspond to qualitatively different phases of the agent's learning trajectory.

Macro-state 0: direct attacking policy. This is the most effective regime. It appears from around $140k$ steps and becomes increasingly frequent later in training. It reaches the highest mean reward (1.256), the highest goal success rate (36.4%), and the shortest episode length (58.6 steps). Its action profile combines `Right` (26.2%), `Release Dribble` (25.9%), `Shot` (18.9%), `Long Pass` (14.2%), and `Bottom-Right` (7.3%). Compared with macro-state 1, the agent is less exclusively driven by moving right and combines forward progression with dribble control, shooting, and occasional long passes. Spatially, this regime is associated with higher offensive-third occupation, a higher average ball position on the attacking axis, and a lower distance to goal. We interpret it as a more efficient vertical attacking policy.

Macro-state 1: unstable offensive transition. This regime contains 26 checkpoints and dominates the intermediate part of training, from approximately $40k$ to $430k$ steps. It improves over macro-state 2, with a mean reward of 0.917 and a goal success rate of 8.3% but remains unstable. The dominant action is `Right` (37.2%), which is meaningful because the controlled team attacks towards the right side of the pitch. The agent also uses `Shot` (17.7%) and `Long Pass` (16.6%), suggesting an emerging offensive intent. However, frequent possession losses and regains show that the policy is still fragile: the agent acts more but does not yet act reliably.

Macro-state 2: sterile possession. This regime corresponds to the earliest stage of training, with 5 checkpoints mostly between $10k$ and $120k$ steps. It has the lowest mean reward (0.708), no scored goals, and the longest episodes (263.7 steps on average). The agent often keeps the ball alive but without converting possession into progress or goals. Its action profile is dominated by `Release Dribble` (31.2%), `Shot` (23.0%), `Idle` (21.2%), and `Release Direction` (20.4%). This indicates a degenerate early policy: the agent shoots often but never scores, remains inactive for many steps, and does not yet use movement or passing in a coherent way. We interpret this regime as sterile possession.

Action-level and pass-network interpretation. The action distributions clarify the tactical transition. In macro-state 2, the high rates of `Idle`, `Release Dribble`, and `Release Direction` suggest that the policy has not yet learned coherent football primitives; the high `Shot` frequency is misleading because it is associated with zero scoring success. In macro-state 1, the emergence of `Right` marks the discovery of the attacking direction but the high number of possession changes indicates that the behaviour remains unstable. In macro-state 0, the agent combines forward movement, dribble release, shooting, and occasional long passing in a more effective way. The resulting behaviour is therefore not a slow possession-based construction but a vertical attacking policy.

The graph-based pass features were introduced to detect collective passing behaviour, following the motivation of Scott et al. [12]. In this run, however, the pass-network indicators remain weak: in macro-state 0, the pass success proxy, pass-network density, and average clustering are essentially zero, and the number of unique ball carriers remains close to 2. Thus, even though `Long Pass` appears in the action distribution, it does not translate into a rich or reliable passing

network. The learned policy should therefore be interpreted as a direct attacking strategy rather than the emergence of sophisticated collective football.

5 Discussion & Limitations

Interpretation and scope. The results suggest that the agent's learning trajectory is not only a smooth reward improvement but can be decomposed into persistent behavioural regimes. The MCA pipeline identifies a progression from sterile possession, to an unstable offensive transition, and finally to a more efficient vertical attacking policy. However, this final policy should not be over-interpreted as sophisticated football: the agent remains mostly direct and weakly collective. The learned strategy is better described as a vertical attacking shortcut than as an emergent collective tactic.

Scenario and computational limitations. The main empirical limitation is the simplicity of the selected scenario. `academy_pass_and_shoot_with_keeper` is useful as a controlled proof of concept but it is not ideal for studying coordinated passing, spatial occupation or team-level tactics. This partly explains why the pass-network metrics remain weak: the agent can succeed through a direct attacking policy rather than a distributed passing strategy.This limitation is also computational. The full pipeline requires training a PPO agent, storing many checkpoints, evaluating each one over several episodes, extracting features and estimating transition matrices. Even this simple scenario took several hours on our available hardware. Richer environments such as `academy_3_vs_1_with_keeper`, `academy_counter_attack_easy` or `5_vs_5` would be more relevant tactically but require substantially more compute.

Methodological limitations. Several behavioural features are proxies rather than exact football annotations. Pass success and pass-network edges, for instance, are reconstructed from GRF observations by detecting a pass action followed by a change of ball carrier within the same team. Spatial features such as width, compactness or distance to goal are also useful summaries but do not fully capture tactical context. The aggregation itself is empirical: micro-states and macro-states are obtained through clustering, and their number remains partly heuristic.

Future work. Future work should repeat the pipeline across several random seeds, reward settings and scenarios of increasing complexity. `academy_3_vs_1_with_keeper` would test richer attacking choices, `academy_counterattack_easy` would introduce a more collective transition situation, and `5_vs_5` would be a stronger test of whether MCA can reveal genuinely collective tactical regimes. Overall, this study should be seen as a proof of concept: MCA can turn a sequence of RL checkpoints into an interpretable dynamical model of behavioural evolution.

6 Conclusion

In this work, we proposed a first application of Markov Chain Aggregation to the study of strategy evolution in a football reinforcement learning agent. Instead of analysing training only through reward curves or isolated descriptive statistics, we represented each checkpoint by a behavioural feature vector, discretised the training trajectory into micro-states, and aggregated them into interpretable macro-states.

On the `academy_pass_and_shoot_with_keeper` scenario, the method reveals a structured progression: the agent moves from sterile possession, to an unstable offensive transition, and finally to a more efficient vertical attacking policy. The *lumpability* test does not reject the resulting macro-chain, suggesting that the aggregation provides a plausible Markovian coarse-graining of the observed behavioural dynamics.

The experiment remains a controlled proof of concept. The final policy is more effective but not genuinely collective: pass-network metrics remain weak and the learned behaviour is best interpreted as a direct attacking shortcut. Nevertheless, the results show that MCA can provide a useful dynamical lens on reinforcement learning training trajectories. By turning a sequence of checkpoints into a small number of behavioural regimes and transition probabilities, the approach opens a path toward more systematic analysis of emergent strategies in richer football environments.

Acknowledgments

The authors thank Julien Randon-Furling for teaching the Interactions course of the MVA Master's programme at ENS Paris-Saclay, for which this work was produced. We also acknowledge Google Research for open-sourcing the Google Research Football environment and the contributors to the Gameplay Football engine on which it is built.

References

[1] Sven Banisch. 2015. *Markov Chain Aggregation for Agent-Based Models.* Springer, Cham.

[2] Lasse Espeholt et al. 2018. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *Proceedings of ICML.*

[3] Ian Graham. 2024. *How to Win the Premier League.* Century, London.

[4] Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. *TrueSkill: A Bayesian Skill Rating System.* Technical Report. Microsoft Research.

[5] Dan Horgan et al. 2018. Distributed Prioritized Experience Replay. In *Proceedings of ICLR.*

[6] Robert W. Jernigan and Robert H. Baran. 2003. Testing lumpability in Markov chains. *Statistics & Probability Letters* 64, 1 (2003), 17–23. doi:10.1016/S0167-7152(03)00126-3

[7] John G. Kemeny and J. Laurie Snell. 1960. *Finite Markov Chains.* Van Nostrand Reinhold, New York.

[8] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. 2020. Google Research Football: A Novel Reinforcement Learning Environment. arXiv:1907.11180 [cs.LG] https://arxiv.org/abs/1907.11180

[9] OpenAI et al. 2019. *Dota 2 with Large Scale Deep Reinforcement Learning.* Technical Report. OpenAI. https://arxiv.org/abs/1912.06680.

[10] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. http://jmlr.org/papers/v22/20-1364.html

[11] John Schulman et al. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[12] Adam Scott, Keisuke Fujii, and Masahiro Onishi. 2022. How does AI play football?. In *Proceedings of ICAART*. SciTePress, Vienna.

[13] David Silver et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529 (2016), 484–489.

A PPO Hyperparameters

Table 1. PPO [11] Training Hyperparameters

Hyperparameter	**Value**
Policy	MlpPolicy
Network architecture	[512, 512]
Total timesteps	500k
Parallel environments	8
Steps per update	512
Batch size	1024
Epochs per update	15
Learning rate	3×10^{-4}
Discount factor γ	0.99
GAE λ	0.95
Clip range ε	0.2
Entropy coefficient	0.01
Value function coefficient	0.5
Max gradient norm	0.5

B Macro-State Behavioural Statistics

Metric	**Macro 0**	**Macro 1**	**Macro 2**
Checkpoints	19	26	5
Training steps	$140k$–$500k$	$40k$–$430k$	$10k$–$120k$
Mean reward	1.256	0.917	0.708
Success rate	36.4%	8.3%	0.0%
Episode length	58.6	97.3	263.7
Passes per episode	7.93	18.25	3.90
Shots per episode	9.49	8.60	58.30

Table 2. Average behavioural statistics by macro-state. The success rate is the fraction of evaluation episodes in which the agent scores.