Ranking futbol with higher order networks

Andres Chacoma*

Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales,

Departamento de Física. Buenos Aires, Argentina and

CONICET - Universidad de Buenos Aires,

Instituto de Física Interdisciplinaria y Aplicada. Buenos Aires, Argentina.

Juan I. Perotti and Orlando V. Billoni

Universidad Nacional de Córdoba, Facultad de Matemática,

Astronomía, Física y Computación. Córdoba, Argentina and

CONICET- Universidad Nacional de Córdoba,

Instituto de Física Enrique Gaviola. Córdoba, Argentina.

Abstract

We propose a unified methodological framework to quantify team performance in elite football by combining event-level performance metrics, higher-order network representations, and algebraic ranking methods. Using data from the 2017-2018 season of the five major European leagues, we construct metric-specific weighted graphs in which teams are connected through relative performance indicators. These graphs are analyzed via Hodge decomposition, and the gradient component is used to derive metric-based team ratings. The resulting rankings are systematically compared with the true league standings using Pearson and Kendall correlation measures, revealing strong metric- and league-dependent effects. While some competitions exhibit high correspondence between individual metric-based ratings and final standings, others display markedly weaker agreement, suggesting the presence of distinct competitive dynamics. To address this heterogeneity, we introduce a composite rating obtained as a parsimonious linear combination of metric-based ratings, optimized separately for each league. This composite approach significantly improves predictive power and allows the relative importance of different performance indicators to be quantified in a league-specific manner. Our results demonstrate how higher-order network methods provide a flexible and interpretable framework to uncover latent performance structures in football, offering a complementary perspective to outcome-based rankings and a general approach applicable to other oppositional sports.

I. INTRODUCTION

In recent years, the study of sports competitions has undergone a significant transformation through the adoption of theoretical frameworks drawn from complexity science [1–10]. Sports championships, characterized by the dynamic interaction of multiple agents (players, teams, contexts), nonlinear outcomes, and the emergence of non-trivial collective patterns, are increasingly understood as complex adaptive systems [11]. This perspective makes it possible to move beyond purely descriptive statistical analyses and to address fundamental questions regarding competitive dynamics, performance evolution, and the underlying structure of play from robust theoretical principles.

* achacoma@df.uba.ar

This intersection between complex systems and sport is fertile from a dual perspective. From an academic standpoint, it offers a unique laboratory for validating and developing theories of networks, competitive dynamics, diffusion processes, and the emergence of hierarchies in social systems governed by clearly defined rules. From the high-performance perspective, the quantitative and objective modeling of competition is currently transforming decision-making processes [12]. Providing coaches, analysts, and decision-makers with data-driven tools to assess performance beyond immediate outcomes, diagnose structural strengths and weaknesses, and optimize strategies has become a priority in the pursuit of sustainable competitive advantages.

While this approach has been successfully applied to various team and individual sports, football stands out as a particularly rich domain. Its global popularity, the growing availability of high-resolution data (event data, tracking data), and the inherent tactical complexity of a continuous-flow, low-scoring game make it an ideal case study. Recent research has employed concepts from complex systems to model ball possession as a diffusion process, characterize collective creativity, and identify critical phases of play, demonstrating the considerable potential of this theoretical framework [13–19].

Within the methodological toolkit of complex systems, complex network theory has emerged as a particularly powerful approach for football analysis. By modeling players or teams as nodes and their interactions (passes, duels, or competitive comparisons) as edges, it becomes possible to quantify and visualize the structural organization of the game. This approach has enabled, for instance, the measurement of player centrality in passing networks, the identification of team-specific tactical patterns, and the assessment of a playing system's robustness to disruptions such as injuries, providing insights that are not accessible through traditional analytical methods [20–26].

In this work, we propose a specific application of network theory to address a central problem in sports analytics: the objective and multidimensional comparison of team performance. Our approach consists of defining a rating system that is not based solely on final match outcomes, but on a set of performance metrics derived from event data. To this end, we construct one network per metric, where nodes represent teams and weighted edges encode their relative performance in direct matchups. Applying the algebraic graph decomposition method *HodgeRank* [27] to these networks, allows us to infer a scalar rating for each team in each dimension of the game, thereby offering a nuanced, process-based comparison.

The specific metrics that feed this model are defined and justified in the following section.

The construction of these multidimensional, process-based ratings has direct practical implications. A single ranking, such as the league table, summarizes outcomes but can obscure the mechanisms underlying them. By decomposing performance into interpretable components (e.g., pressure, chance creation, possession times), our framework provides a structural diagnosis. For coaching staffs, this makes it possible to identify teams that accumulate points despite poor performance in key metrics—an indication of potential unsustainability—or, conversely, teams whose strong process-level performance does not translate into results, suggesting inefficiency in decisive moments. Such information is crucial for informed decision-making in areas such as player recruitment, tactical design, and opponent-specific preparation, allowing resources to be directed toward the aspects of play that truly determine long-term competitive advantage.

This article is organized as follows to present the methodological framework and findings in a coherent manner. Section II, Data and Metrics, describes in detail the data collection process and sources, defines the eight performance metrics derived from match events, and explains the construction of the complex networks that represent competitive interactions between teams for each metric. Section III, Theoretical Framework, presents the algebraic foundations of the HodgeRank method, detailing how this graph decomposition technique is applied to infer ratings from the pairwise comparisons encoded in our networks. The main findings are presented in Section IV, Results, structured into four analytical subsections: (IV A) an initial statistical analysis of the true rating and the validation of the hierarchical skill model; (IV B) a global correlation analysis (Pearson coefficient) between the metric-based ratings and the true rating; (IV C) a comparison of the resulting ordinal rankings using Kendall's rank correlation coefficient; and (IV D) the proposal and optimization of a composite rating, obtained as a weighted linear combination of the most relevant metric ratings. Finally, the article concludes with Section V, Discussion and Conclusions, which integrates and interprets the overall results, discusses their implications, outlines the limitations of the study, and proposes concrete directions for future research.

II. DATA AND METRICS

A. Metric collection

In this work, we use the event data set provided by L. Pappalardo et al. in [28]. In that article, the authors manually annotate all matches from the 2017–2018 season of the major European football leagues: La Liga (Spain), Premier League (England), Serie A (Italy), Bundesliga (Germany), and Ligue 1 (France). For each match, they detect, classify, and localize in time and space all relevant events, including goals, shots, passes, corner kicks, fouls, among others. In the reference system used to locate events in time and space, t denotes the elapsed time since the start of the match, the coordinate x represents the distance relative to the goal defended by the team generating the event, and the coordinate y represents the distance relative to the right touchline. Spatial coordinates are expressed as percentages of the pitch length, such that, for example, $x = 0$, $x = 50$, and $x = 100$ correspond to the team's own goal line, the midfield line, and the opponent's goal line, respectively.

Within this framework, we define a ball possession interval (BPI) as the set given by a continuous sequence of events generated by a single team. Note that each BPI contains information from only one team. We collect all BPIs from all teams in each league, and from these data we extract metrics that allow us to detect some of the tactical resources employed by teams within that temporal window of the match. The metrics considered in our analysis are based on those proposed by J. Fernández-Navarro in [29]. The study of these metrics was also employed in a previous work to identify characteristic football playing styles [18]. Below, we describe each of these metrics in detail:

1. *Direct play.* Each time a pass or a free kick occurs within a BPI, we measure the average velocity in the attacking direction, defined as the ratio between the distance traveled by the ball along the x axis and the elapsed time. For each BPI, we retain the maximum value. This metric allows us to quantify how directly the team moves the ball toward the opponent's goal.

2. *Counterattack.* Given two consecutive events within a BPI, if the first event occurs at $x_1 < 40$ and the second at $x_2 > 60$ with a temporal difference Δt, the velocity is reported as $v = \frac{x_2 - x_1}{\Delta t}$. Otherwise, a value of 0 is reported. This metric measures

how quickly a team transitions from a defensive position in its own half to an offensive position in the opponent's half.

3. *Build up.* If within a BPI it is verified that $\bar{x} > 60$, i.e., the possession develops predominantly in the opponent's half, the total possession time is reported. Otherwise, a value of 0 is reported. This metric captures possession time in situations where the team strongly occupies the opponent's territory.

4. *Midfield play.* If within a BPI it is observed that $\bar{x} \leq 60$ and $\bar{x} \geq 40$, i.e., the possession develops predominantly in the central area of the pitch, the total possession time is reported; otherwise, a value of 0 is reported. The purpose of this variable is to measure the time the team spends in the midfield zone.

5. *Flow rate.* For each BPI satisfying $\bar{x} \geq 50$, we compute the temporal differences between all consecutive events and calculate their mean value, $\bar{dt}$. The metric is then defined as $1/\bar{dt}$. In this way, this metric provides a measure of how quickly the team circulates the ball in the opponent's half.

6. *Crossing.* If a crossing event is observed within a BPI, a value of 1 is reported; otherwise, a value of 0 is reported. This metric is used to count attempts to reach the penalty area through aerial play.

7. *Pressure point.* For each BPI, we take the first event and extract its x coordinate, i.e., the position at which the team starts its possession. This allows us to assess whether the team is recovering the ball in its own half, the midfield zone, or the opponent's half.

8. *Shots.* If a "Shot" event is recorded within a BPI, a value of 1 is reported; otherwise, a value of 0 is reported. This metric allows us to count the number of shots on goal produced by each team.

For our analysis, all BPIs with fewer than 3 events and with a total duration shorter than 2 seconds were discarded. The rationale behind this choice is to exclude brief, transient recoveries and retain only consolidated possessions. From the data collection process, a total of 215 681 BPIs were obtained. After computing the metric values for each BPI, we analyzed their empirical distributions. We observed that the metrics approximately follow log-normal

distributions; therefore, we applied the transformation $x \to \log(1 + x)$ in order to work with distributions that are closer to normality. Subsequently, we aggregated the information by match and by team, summing the values obtained for each metric. In this way, for example, the feature *Shots* quantifies the total number of shots on goal taken by a team in a given match. Similarly, the feature *Build up* quantifies the net amount of time during which a team sustained an attacking possession against its opponent in that match. Note that in the 2017–2018 season, teams in the Spanish, English, French, and Italian leagues played 38 matches. Therefore, when considering data from the top 4 teams in each league, each league contributes a total of $38 \times 4 = 152$ samples to the dataset. In the German league, due to the smaller number of teams, 34 matches were played, and thus this league contributes 136 samples. Consequently, the data matrix consists of 744 rows and 10 columns. Finally, in a separate dataset we collected meta-data associated with each sample, which are later used in the analysis: the team corresponding to each sample, the league it belongs to, and the final position in the league table.

B. Representation as complex networks

In what follows, we present our proposal to represent performance metrics in terms of complex networks. We define $M(i,\, j,\, g)$ as the performance metric corresponding to team i when facing team j in match g. For example, this quantity may represent the number of shots on goal taken by FC Barcelona when playing against Real Madrid in the first encounter of the Spanish league *La Liga*. In our dataset, all teams participated in a round-robin league format, facing each opponent twice: a first-leg match (g_1) and a second-leg match (g_2). Using the information from both encounters, we define an aggregated metric that summarizes the performance observed between the two teams over the course of the season:

$$M(i,\, j) = \sum_{g_1, g_2} M(i,\, j,\, g).$$

In the previous example, $M(i,\, j)$ represents the total number of shots on goal taken by FC Barcelona against Real Madrid across both matches of the tournament. By computing $M(i,\, j)$ for each pair of teams in a league L, it is possible to represent these performance relationships by means of a directed and weighted graph $G(L,\, M)$, whose weights are defined as

$$f_{ij} = M(j,\, i) - M(i,\, j).$$

Note that, in this representation, $f_{ij} < 0$ indicates that team i outperformed team j with respect to the metric under consideration. Within this framework, we construct a total of 50 graphs, corresponding to 8 graphs per league, each one associated with a different performance metric.

III. THEORETICAL FRAMEWORK

IV. RESULTADOS

A. Statistics of the true rating

The purpose of this section is to define a statistical model for the probability distribution of the true rating, R_T, defined as the total number of points obtained by a team over the course of a league season. R_T is a stochastic variable that depends on the number of matches won, drawn, and lost by teams in the league. In football leagues, a team is awarded 3 points for a win, 1 point for a draw, and 0 points for a loss. Within this framework, R_T can be expressed in terms of the stochastic variables W and D, which represent the number of matches won and drawn, respectively,

$$R_T = 3W + D.$$

Let n denote the total number of matches in the season. We model W using a binomial distribution,

$$W \sim Bin(n|p_w),$$

where p_w is the probability of winning a match. Conditional on the number of wins $W = w$, the number of draws follows,

$$D \mid W = w \sim Bin\left(n - w \;\middle|\; \frac{p_d}{1 - p_w}\right),$$

where p_d is the probability of drawing a match. Since teams exhibit different competitive levels, each team is characterized by intrinsic probabilities p_w and p_d of winning and drawing, respectively. We model these probabilities using a hierarchical approach: each team is

endowed with a pair of latent skills $\boldsymbol{\eta} = (\eta_w,\ \eta_d)^T$, corresponding to win and draw abilities, which follow a bivariate normal distribution,

$$\boldsymbol{\eta} \sim \mathcal{N}\ (\boldsymbol{\mu},\ \boldsymbol{\Sigma})\ ,$$
$$\boldsymbol{\mu} = (\mu_w,\ \mu_d)^T, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_w^2 & \rho\sigma_w\sigma_d \\ \rho\sigma_w\sigma_d & \sigma_d^2 \end{pmatrix}\ ,$$

where $\boldsymbol{\mu}$ is the center of the distribution and $\boldsymbol{\Sigma}$ is the covariance matrix. In this framework, ρ represents the correlation between win and draw abilities. The probabilities are obtained through a softmax transformation,

$$p_w = \frac{e^{\eta_w}}{1\ +\ e^{\eta_w}\ +\ e^{\eta_d}}, \quad p_d = \frac{e^{\eta_d}}{1\ +\ e^{\eta_w}\ +\ e^{\eta_d}}.$$

This transformation ensures that $p_w + p_d \leq 1$ and that the model is identifiable by implicitly fixing $\eta_l = 0$ for losses. With these elements, we can write the joint distribution of wins and draws by marginalizing over the random effects $\boldsymbol{\eta}$,

$$P(W = w,\ D = d) = \int_{\mathbb{R}^2} Bin(n|p_w(\boldsymbol{\eta}))\ Bin\left(n\ -\ w\ \middle|\ \frac{p_d(\boldsymbol{\eta})}{1\ -\ p_w(\boldsymbol{\eta})}\right)\ \phi(\boldsymbol{\eta}\ |\ \boldsymbol{\mu},\ \boldsymbol{\Sigma})\ d\boldsymbol{\eta}, \tag{1}$$

Note that the integral in Eq. 1 does not admit a closed-form solution; therefore, we evaluate it using Monte Carlo methods. With these ingredients, we can finally write the theoretical probability distribution of the true rating as

$$P(R_T = r_T) = \sum_X P(W = w,\ D = d), \tag{2}$$
$$X = \{w,\ d > 0,\ \ 3w\ +\ d = r_T,\ \ w\ +\ d \leq n\}.$$

The degrees of freedom provided by the parameters μ_w, μ_d, σ_w, σ_d, and ρ allow the theoretical curve in Eq. 2 to be fitted to the empirical distribution. The fit was performed using the dataset associated with the English, French, Italian, and Spanish leagues. For this analysis, we chose not to include data from the German league, as its competition involves a smaller number of teams (18), which leads to a slightly different distribution of true rating values compared to the other leagues, where a total of 20 teams participate. The fitting procedure was carried out using the Nelder–Mead algorithm, minimizing the RMSE between the empirical and theoretical CDFs. Subsequently, using the optimal parameters, we performed parametric bootstrapping (1000 replicas) to estimate the mean values and

uncertainties of the probabilities p_w and p_d. For the probability of winning, we obtained $\bar{p}_w = 0.366$ ($SD = 0.010$) with a 95% confidence interval of [0.3527, 0.3798]. Similarly, for the probability of drawing, we obtained $\bar{p}_d = 0.280$ ($SD = 0.020$) with a 95% confidence interval of [0.2548, 0.3074]. To contrast these results with empirical data, we computed for each team i the empirical probabilities of winning and drawing, q_w and q_d, using match outcome information,

$$q_w^{(i)} = \frac{\#\ wins^{(i)}}{\#\ matches^{(i)}}, \quad q_d^{(i)} = \frac{\#\ draws^{(i)}}{\#\ matches^{(i)}}.$$

By computing the mean value and standard deviation over the set of all teams, we obtained for the probability of winning $\bar{q}_w = 0.379$ ($SD = 0.168$) with a 95% confidence interval of [0.1572, 0.7638], and for the probability of drawing $\bar{q}_d = 0.243$ ($SD = 0.081$) with a 95% confidence interval of [0.1046, 0.3954]. From these results, we observe the following: (i) the mean values are similar (0.366 vs. 0.379 for wins, 0.280 vs. 0.243 for draws), which validates the center of the distribution; (ii) the empirical standard deviations (0.168, 0.081) are substantially larger than those implied by the model (0.01, 0.02), indicating that the model underestimates team-level heterogeneity. Regarding the correlation parameter, the parametric bootstrap yielded $\bar{\rho} = -0.130$ ($SD = 0.07$) with a 95% confidence interval of $[-0.3040, -0.0371]$. This result indicates the presence of a weak negative structural correlation between teams' win ability η_w and draw ability η_d. This suggests that teams with a higher propensity to win tend, on average, to exhibit a slightly lower propensity to draw. In Fig. 1(a), we show the CDF of the true rating together with the fitted model. The proposed model accurately captures the behavior of the empirical curve across the entire support. In Fig. 1(b), we present a quantile–quantile plot comparing theoretical and empirical quantiles. For reference, we also include the relationship between the data quantiles and those associated with a Gaussian distribution with mean and standard deviation equal to the sample values, $\bar{R}_T = 52.45$ ($SD = 18.56$). First, we observe that the proposed model reproduces well the behavior in the central-left region and provides an acceptable description of the right tail. The comparison with the Gaussian model reveals that the empirical distribution exhibits lighter left tails and heavier right tails relative to a Gaussian distribution, thereby uncovering the presence of asymmetry. The downward concavity observed in the central region of the plot further indicates that the median of the data is shifted to the left.

B. Correlation between true and metric-based ratings

In this section, we characterize the differences between the true rating, as given by the league point system, and the rating obtained from the HodgeRank method, which we refer to as the metric rating. Since eight performance metrics are considered (see Section II A), for each team we compute eight different metric rating values, which are then compared with the true rating. It is important to emphasize that the true rating and the metric rating are expressed on different scales. The former is measured in terms of total points obtained, whereas the latter is expressed as a standardized score. Nevertheless, this does not prevent a meaningful comparison, as the relevant information conveyed by these quantities is not their absolute value, but rather the relative differences between teams. For this reason, in order to compare the behavior of the true rating and the metric rating, we standardize both variables and work with their corresponding *z-scores*.

In Fig. 2(a), we show the CDF of the true rating computed over all teams and all leagues, together with the CDF of the metric rating computed over all teams, all leagues, and all eight metrics considered. For both curves, outlier values—defined as those above the quantile $Q_{99.7}$—were removed. We observe a remarkable agreement across most of the range, with some differences appearing in the tails: the metric rating exhibits more extreme positive values relative to the mean, whereas the true rating shows slightly more extreme negative values. In Fig. 2(b), we show the distribution of metric rating values for each league. For this purpose, within each league we aggregate the metric rating values associated with all metrics and standardize them using the z-score. As a reference, the CDF of a standard normal distribution is also shown as a black dashed line. It can be observed that the curves associated with different leagues differ slightly, indicating league-dependent behavior. Moreover, in all cases we observe an accumulation of probability mass in the central region. This is compensated by lighter left tails and heavier right tails relative to the standard normal distribution. On the other hand, in Fig. 2(c) we plot the true rating and the metric rating as a function of the true ranking, for the case of the *Pressure point* metric in the English league. The purpose of this representation is to highlight team-by-team differences. First, we note that the true rating decreases as the ranking increases. This decrease is not strictly monotonic, as plateaus may occur, but it is non-increasing by construction: the first-ranked team has a true rating greater than or equal to that of the second-ranked team, the

second greater than or equal to the third, and so on until the last position. For the metric rating curve, we also observe an overall decreasing trend, although not strictly. In qualitative terms, the metric rating follows the general trend of the true rating, while exhibiting small local deviations. In Fig. 2(d), we show an analogous comparison for the *Direct play* metric in the Spanish league. In this case, the metric rating curve does not display a clear decreasing trend and shows a strong decorrelation with respect to the true rating values. These results indicate that, for a given league, the true rating may differ substantially or only marginally from the metric rating, depending on the specific metric used to infer the rating.

Our next objective is to quantify the correlation between the true rating curve of a given league and the metric rating curves associated with each performance metric. Specifically, we aim to identify which metric yields a metric rating that best follows the trend of the true rating, and whether this relationship depends on the league. To this end, we compute the Pearson correlation coefficient between the true rating of each league and the metric rating obtained from each metric. The results are shown in Fig. **??**. Each panel corresponds to a different metric, and within each panel each bar represents a league. The panels are ordered from highest to lowest average correlation, and the bars within each panel are ordered from highest to lowest correlation value. Hatched bars indicate cases for which statistical significance cannot be established. First, we observe that the Italian and English leagues exhibit the highest correlations across all metrics. Second, we find that the French and German leagues alternate between third and fourth place among cases with intermediate correlation values for almost all metrics. Finally, we observe that the Spanish league presents the weakest results, showing the lowest correlation values in 7 out of the 8 metrics considered. Nevertheless, it still exhibits strong correlations ($\rho > 0.4$) for the *Flow rate* and *Counterattack* metrics.

C. Correlation between true and metric-based rankings

In this section, we complement the analysis presented in the previous section by studying the correlation between rankings. We define the *true ranking* of a team as its final position in the league standings. This ranking is obtained by ordering the teams in descending order according to the total number of points accumulated; in the event of a tie, the final ranking is determined by the goal difference (goals scored minus goals conceded). Similarly, we define

the *metric ranking* by ordering the teams in a league in descending order according to the value of the *metric rating*. In this case, no ties were observed during the analysis. To quantify the correlation between the two rankings, we employ Kendall's rank correlation coefficient [30], which is based on a nonparametric hypothesis test whose statistic, $\tau \in [-1, 1]$, takes values close to one when the observations exhibit a similar ordering, and values close to zero when the orderings differ substantially. Figure 4 presents the corresponding results. Analogously to Fig. **??**, each panel corresponds to a metric, while each bar within a panel represents a league. The panels are ordered from highest to lowest according to the average Kendall coefficient, and the bars within each panel are ordered in decreasing order according to their individual values. Bars with hatching indicate cases for which statistical significance cannot be ensured. First, we observe that the Italian league consistently leads all panels, exhibiting the highest values of Kendall's coefficient. The English league generally ranks second, although it is surpassed by the French league in two of the metrics. By contrast, the German and Spanish leagues tend to occupy the lowest positions. This behavior is, in principle, consistent with the results previously observed for the *ratings*. Second, we find that Kendall's coefficients are, in general, slightly lower than the corresponding Pearson correlation coefficients. This difference arises from the nonparametric nature of Kendall's coefficient. Importantly, this discrepancy does not imply a weak correlation, but rather reflects the increased sensitivity of rank-based measures to small fluctuations in regions of the table where point differences between teams are minimal. While Pearson's coefficient validates the global hierarchy of the league, Kendall's coefficient highlights that the exact ordering of teams can be affected by small variations (noise), particularly in regions of the standings with a high density of points, that is, where multiple teams are tied or nearly tied.

D. Composite rating based on performance metrics

In this subsection, we develop a Composite Rating (CR) for each league, defined as a linear combination of the metric ratings obtained from performance metrics. The aim is to provide a more comprehensive and general assessment of team quality.

Let n denote the number of teams in a league and m the number of performance metrics. We define $R^{n\times m}$ as the matrix whose rows contain the metric rating values obtained for each team. Similarly, we define $T^{n\times 1}$ as the vector containing the true ratings of each

team in the league. Within this framework, our objective is to determine the weight vector $\alpha^{m\times 1} = (\alpha_1, \ldots, \alpha_m)^T$ such that

$$T = \alpha R.$$

To solve this system while ensuring both parsimony and explanatory power of the composite rating, we employ an Exhaustive Feature Selection (EFS) algorithm. After normalizing the metrics using standard scaling (z-score), this procedure evaluates all possible combinations of the eight initial performance metrics, ranging from univariate models to the full model including all eight variables.

To prevent overfitting, we adopt the Bayesian Information Criterion (BIC) as the model selection criterion. The BIC identifies the model that maximizes the likelihood of the data while introducing a penalty proportional to model complexity (i.e., the number of variables), thus favoring the most efficient structure. Within this framework, the optimal model is defined as the one that minimizes the BIC value, ensuring that the resulting composite rating is statistically robust and representative of true team performance.

Finally, to compute the composite rating for each league, we fit an Ordinary Least Squares (OLS) regression model using the metrics selected by the EFS algorithm. Note that, since all variables were previously standardized, the estimated coefficients directly reflect the relative importance of each performance dimension in predicting the number of points obtained in the league.

Table I reports the regression results. The analysis of the coefficients highlights key areas for performance improvement.

For the English league, the most relevant metrics, ranked by relative importance, are *Build up*, *Flow rate*, and *Pressure point*, which together explain nearly 90% of the variance in the data. The negative coefficient associated with *Build up* (-24.56) suggests that teams that reduce their build-up time in the final third of the pitch (see the metric definition in Section II A) significantly increase their probability of earning points. Conversely, an increase in this metric penalizes the composite rating, indicating inefficiency in tactical styles that rely heavily on prolonged build-up play in this league. In contrast, the positive coefficient for *Flow rate* (22.69) indicates that teams increasing the speed at which they move the ball across the pitch improve their chances of accumulating points. Similarly, the positive coefficient for *Pressure point* suggests that teams applying higher pressure up the field are more likely to obtain points.

For the French league, a similar coefficient-based analysis reveals that teams that apply high pressure and reduce direct play tend to increase their likelihood of earning points.

The German league presents perhaps the most intriguing case. Notably, a single variable—*Direct play*—is sufficient to explain 46% of the variance. In other words, understanding this metric alone provides substantial insight into why teams gain or lose points in this league, resulting in an extremely parsimonious model.

The results for the Italian league suggest that success is associated with increases in *Flow rate* and *Direct play*, along with a decrease in *Midfield play*. That is, the most successful teams in this league appear to favor a playing style characterized by rapid ball circulation and reduced control time in the midfield.

Finally, in the Spanish league, point accumulation appears to be associated with lower levels of high pressing, combined with increases in ball circulation speed and counterattacking play.

The composite rating obtained for each league can be compared with the true rating in the same manner as was done in the previous sections for individual metric-based ratings. In Fig. 5, panels (a), (b), (c), (d), and (e) show league-by-league, team-by-team comparisons in plots of rating versus true ranking.

In the case of the English league (Fig. 5(a)), we observe that the top two teams exhibit true ratings higher than their composite ratings. These teams therefore display statistical overperformance, appearing to obtain better results than expected based on their underlying metrics. This may be attributable to external factors or stochastic effects. In contrast, the team ranked third shows a true rating lower than its composite rating, indicating statistical underperformance—suggesting that the team performed adequately in process terms but failed in execution or was affected by unfavorable randomness. Similar patterns are observed across all leagues and at all positions in the table.

Overall, the composite rating curves follow trends similar to those of the true rating curves, with particularly strong agreement observed for the English and Italian leagues. To conclude the analysis, Fig. 5(f) presents a bar chart showing the values of the Pearson correlation coefficient ρ and the Kendall rank correlation coefficient τ. In all cases, these values exceed those obtained using individual metrics alone (see Fig. **??** and Fig. 4), with especially notable improvements for the Italian league and significant gains for the Spanish and German leagues.

V. DISCUSSION AND CONCLUSIONS

In this work, we present a unified methodological framework that integrates advanced football performance metrics, graph theory, and a robust algebraic method (HodgeRank) to generate team ratings and rankings.

As an initial step of the analysis, and with the aim of establishing a solid baseline for subsequent evaluation, we examined the statistical properties of the true ranking, defined by the total number of points obtained by teams in their respective leagues. To this end, we proposed a hierarchical model in which each team is characterized by three specific latent skill parameters: the abilities to win, draw, and lose. Analogously to rating systems such as the Elo method [31], these skills define a probability distribution over match outcomes (win, draw, loss). Through a standard numerical fitting procedure, we estimated the corresponding theoretical probabilities, which showed satisfactory agreement with the observed empirical frequencies, thereby validating the adequacy of the proposed model. For the generation of metric-based ratings, we collected eight performance metrics from an event-level database corresponding to the 2017/2018 season of the five major European leagues (England, France, Germany, Italy, and Spain). Based on these metrics, we constructed weighted undirected graphs, where nodes represent teams and edge weights encode relative performance information for a given metric. This procedure resulted in eight networks per league. The Hodge decomposition was then applied to each graph, using the gradient component as the mechanism to derive a rating for each team. In this way, eight ratings per league (one per metric) were obtained and subsequently compared with the true ranking. The relationship between these metric-based ratings and the true ranking was analyzed using Pearson's correlation coefficient to assess global trends, complemented by Kendall's rank correlation coefficient to evaluate the similarity in team ordering. The results revealed clear and consistent patterns, highlighting a strong dependence on both the league and the metric considered. In particular, the English and Italian leagues exhibited the highest levels of similarity between ratings derived from individual metrics and the true ranking, suggesting that, in these competitions, the selected metrics are individually effective predictors of overall performance. In contrast, the Spanish league showed the lowest levels of correspondence, with markedly weaker results across most metrics. This indicates that, for this league, the proposed metrics have limited predictive power and that alternative or complementary in-

dicators may be required. These league-dependent differences further point to the existence of distinctive playing styles or competitive dynamics, a phenomenon already documented in the literature [18]. Finally, once the individual ratings were characterized, we defined for each league a composite rating, constructed as a weighted linear combination of the eight metric-based ratings. The weights were optimized through a fitting procedure designed to identify the most parsimonious combination that best approximates the true ranking. This approach allowed us to obtain a more robust composite rating, significantly improving the predictive capability relative to individual metrics, with particularly notable improvements in the Spanish and German leagues. Moreover, this technique enabled us to quantify the relative importance of each metric within the context of each league, thereby providing an integrated view of team performance. It is important to emphasize that this work does not propose a replacement for the official league table, whose competitive value and sporting legitimacy are not being questioned. We acknowledge that the inherently stochastic and unpredictable nature of football, which is often reflected in the official standings, constitutes an essential part of the sport's appeal. Rather, the objective of our work is to provide a complementary analytical tool that, by uncovering the underlying statistical patterns of performance, offers a richer and more objective perspective on the competitive dynamics of a league. Such a tool could be used by coaching staffs to achieve a more comprehensive evaluation of teams' true strength, beyond the contingencies of immediate match outcomes. Regarding direct applications, the proposed methodological framework is generalizable to any head-to-head sport and to any set of performance metrics, thereby overcoming a key limitation of traditional rating models that rely exclusively on match outcomes (wins, draws, losses). By focusing on indicators of overall performance during the competitive process, our approach provides a potentially more stable evaluation with greater medium-term predictive power, as it is less sensitive to the randomness of isolated results. Specifically, the proposed composite rating makes it possible to identify significant discrepancies between on-field performance and the points actually obtained. This capability is of substantial practical value: on the one hand, it can highlight teams with high ratings but low point totals, suggesting solid performance accompanied by bad luck or inefficiency in decisive moments. On the other hand, it can detect teams with low ratings that nonetheless occupy high positions in the standings, indicating performance sustained by favorable random factors. Identifying such statistically under- or over-performing teams provides coaches and analysts with valu-

able information to adjust strategies, manage expectations, and support technical decisions with a more solid quantitative basis. Several avenues for future research emerge from this work. First, it would be essential to further investigate the inter-league differences observed. Incorporating contextual variables—such as financial resources, home advantage, or prevailing tactical models—would allow the development of a robust explanatory model describing how different football cultures are reflected in performance data and, consequently, in the derived ratings. Such contextual analysis would not only help explain observed differences but also inform the development of more adaptive models. In this regard, a second priority direction would be to refine the composite rating through the incorporation of machine learning techniques and nonlinear models. These methods could dynamically optimize metric weights as a function of the specific context of each league or season, thereby overcoming the limitations of fixed-weight approaches. Furthermore, given the model's demonstrated ability to identify teams whose statistical performance diverges from their actual point totals, a natural and practically relevant extension would be the development of an early warning system. By integrating real-time data, such a system could continuously quantify a "luck factor" or efficiency in critical moments, providing coaching staffs with an analytical tool usable throughout the season. Third, it would be of great interest to validate and extend the proposed framework to other team sports with similar confrontation dynamics. Generalization to disciplines such as basketball or hockey, using their specific performance metrics, would test the robustness of the approach and enable comparative inter-sport studies. Finally, a significant methodological opportunity not yet explored in our analysis lies in fully exploiting the Hodge decomposition. Beyond the gradient component used to generate the ratings, the cyclic and harmonic components contain valuable structural information. The cyclic component, which captures inconsistencies in pairwise comparisons, could quantify the intrinsic competitiveness or unpredictability of a league. The harmonic component, inherent to the topology of the graph, offers a complementary structural perspective. Although its contribution is null in a complete league graph, its analytical potential emerges when connectivity is redefined. For instance, if teams are connected not only through direct matchups but also through similarity in their metric profiles (forming a tactical similarity graph), the harmonic component could reveal natural groupings or communities of teams with similar playing styles. This would allow leagues to be characterized not merely as linear hierarchies, but as networks with possible tactical clusters, substantially enriching the

description of competition beyond simple rankings. The quantitative analysis and integration of these components would significantly enhance the characterization of a competition, adding a deeper interpretative layer to traditional classifications.

---

[1] Marc Barthelemy. Fragility of chess positions: Measure, universality, and tipping points. Physical Review E, 111(1):014314, 2025.

[2] A Clauset, M Kogan, and S Redner. Safe leads and lead changes in competitive team sports. Physical Review E, 91(6):062815, 2015.

[3] Javier M Buldú, Javier Busquets, Ignacio Echegoyen, and F Seirul. lo. Defining a historic football team: Using network science to analyze guardiola's fc barcelona. Scientific reports, 9(1):13602, 2019.

[4] Haroldo V Ribeiro, Satyam Mukherjee, and Xiao Han T Zeng. Anomalous diffusion and long-range correlations in the score evolution of the game of cricket. Physical Review E, 86(2):022102, 2012.

[5] Andrés Chacoma and Orlando V Billoni. Simple mechanism rules the dynamics of volleyball. Journal of Physics: Complexity, 3(3):035006, 2022.

[6] Andrés Chacoma and Orlando V Billoni. Probabilistic model for padel games dynamics. Chaos, Solitons & Fractals, 174:113784, 2023.

[7] Andrés Chacoma and Orlando V Billoni. Emergent complexity in the decision-making process of chess players. Scientific Reports, 15(1):23234, 2025.

[8] Chiara Zappalà, Alessandro Pluchino, Andrea Rapisarda, Alessio Emanuele Biondo, and Pawel Sobkowicz. On the role of chance in fencing tournaments: An agent-based approach. Plos one, 17(5):e0267541, 2022.

[9] Sergio J Ibáñez, Aitor Mazo, Juarez Nascimento, and Javier García-Rubio. The relative age effect in under-18 basketball: Effects on performance according to playing position. PloS one, 13(7):e0200408, 2018.

[10] Javier Galeano, Miguel-Ángel Gómez, Fernando Rivas, and Javier M Buldú. Using markov chains to identify player's performance in badminton. Chaos, Solitons & Fractals, 165:112828, 2022.

[11] Pedro Silva, Luís Vilar, Keith Davids, Duarte Araújo, and Júlio Garganta. Sports teams as

complex adaptive systems: manipulating player numbers shapes behaviours during football small-sided games. SpringerPlus, 5(1):191, 2016.

[12] Michael Ashford, Andrew Abraham, and Jamie Poolton. Understanding a player's decision-making process in team sports: a systematic review of empirical evidence. Sports, 9(5):65, 2021.

[13] Johann H Martínez, David Garrido, José L Herrera-Diestra, Javier Busquets, Ricardo Sevilla-Escoboza, and Javier M Buldú. Spatial and temporal entropies in the spanish football league: A network science perspective. Entropy, 22(2):172, 2020.

[14] Xiaoxiang Cao, Xiaodong Zhao, Huan Tang, Nianchun Fan, and Fateh Zereg. Football players' strength training method using image processing based on machine learning. Plos one, 18(6):e0287433, 2023.

[15] Ken Yamamoto, Seiya Uezu, Keiichiro Kagawa, Yoshihiro Yamazaki, and Takuma Narizuka. Theory and data analysis of player and team ball possession time in football. Physical Review E, 109(1):014305, 2024.

[16] A Chacoma, Nahuel Almeira, Juan Ignacio Perotti, and Orlando Vito Billoni. Modeling ball possession dynamics in the game of football. Physical Review E, 102(4):042120, 2020.

[17] A Chacoma, N Almeira, JI Perotti, and OV Billoni. Stochastic model for football's collective dynamics. Physical Review E, 104(2):024110, 2021.

[18] Andres Chacoma and Orlando V Billoni. Data-driven approach to defining football styles in major leagues. Chaos, Solitons & Fractals, 200:116926, 2025.

[19] Guy Amichay, Hugo Silva, João Brito, and Rui Marcelino. Characterizing the spatial structures of competing football teams. Scientific Reports, 15(1):35217, 2025.

[20] Sergio Caicedo-Parada, Carlos Lago-Peñas, and Enrique Ortega-Toro. Passing networks and tactical action in football: A systematic review. International Journal of Environmental Research and Public Health, 17(18):6649, 2020.

[21] Genki Ichinose, Tomohiro Tsuchiya, and Shunsuke Watanabe. Robustness of football passing networks against continuous node and link removals. Chaos, Solitons & Fractals, 147:110973, 2021.

[22] Bruno Gonçalves, Diogo Coutinho, Sara Santos, Carlos Lago-Penas, Sergio Jiménez, and Jaime Sampaio. Exploring team passing networks and player movement dynamics in youth association football. PloS one, 12(1):e0171156, 2017.

[23] A Chacoma, OV Billoni, and MN Kuperman. Complexity emerges in measures of the marking dynamics in football games. Physical Review E, 106(4):044308, 2022.

[24] Andrés Chacoma. Identification and optimization of high-performance passing networks in football. Physical Review E, 111(4):044313, 2025.

[25] Ming-Xia Li, Li-Gong Xu, and Wei-Xing Zhou. Motif analysis and passing behavior in football passing networks. Chaos, Solitons & Fractals, 190:115750, 2025.

[26] Kate KY Yung, Paul PY Wu, Karen aus der Fünten, Anne Hecksteden, and Tim Meyer. Using a bayesian network to classify time to return to sport based on football injury epidemiological data. PloS one, 20(3):e0314184, 2025.

[27] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. Mathematical Programming, 127(1):203–244, 2011.

[28] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. Scientific data, 6(1):236, 2019.

[29] Javier Fernandez-Navarro, Luis Fradua, Asier Zubillaga, and Allistair P. McRobert. Influence of contextual variables on styles of play in soccer. International Journal of Performance Analysis in Sport, 18(3):423–436, 2018.

[30] Maurice G Kendall. A new measure of rank correlation. Biometrika, 30(1-2):81–93, 1938.

[31] Arpad E. Elo. The Rating of Chess Players, Past and Present. Arco Publishing, New York, 1978.

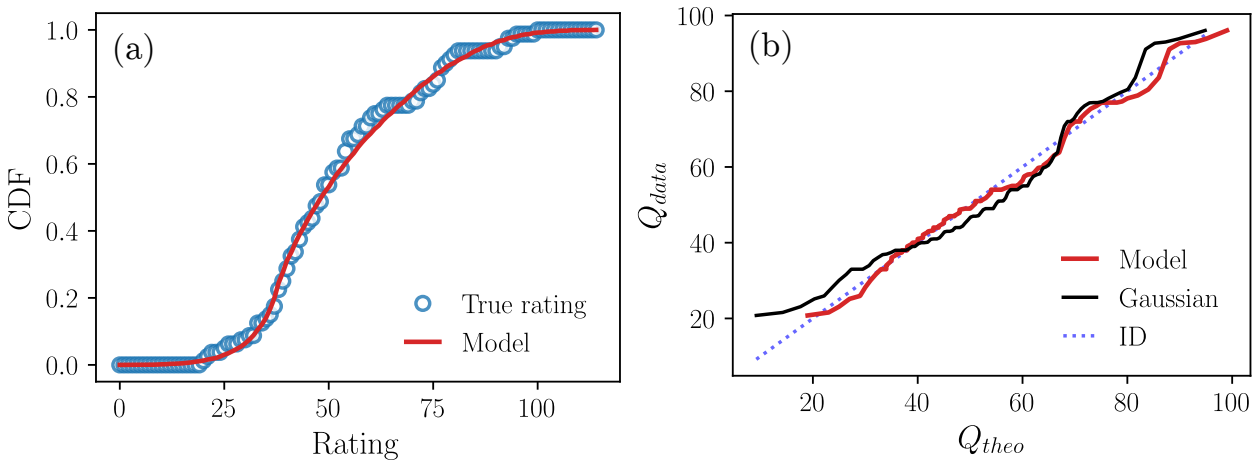

FIG. 1. Statistics of the true rating. (a) Comparison between the cumulative distribution function (CDF) of the true rating values and the distribution obtained from the proposed model. (b) Relationship between the empirical quantiles and those obtained from the model. For reference, a comparison with a Gaussian distribution with mean and standard deviation equal to those of the data is also shown.

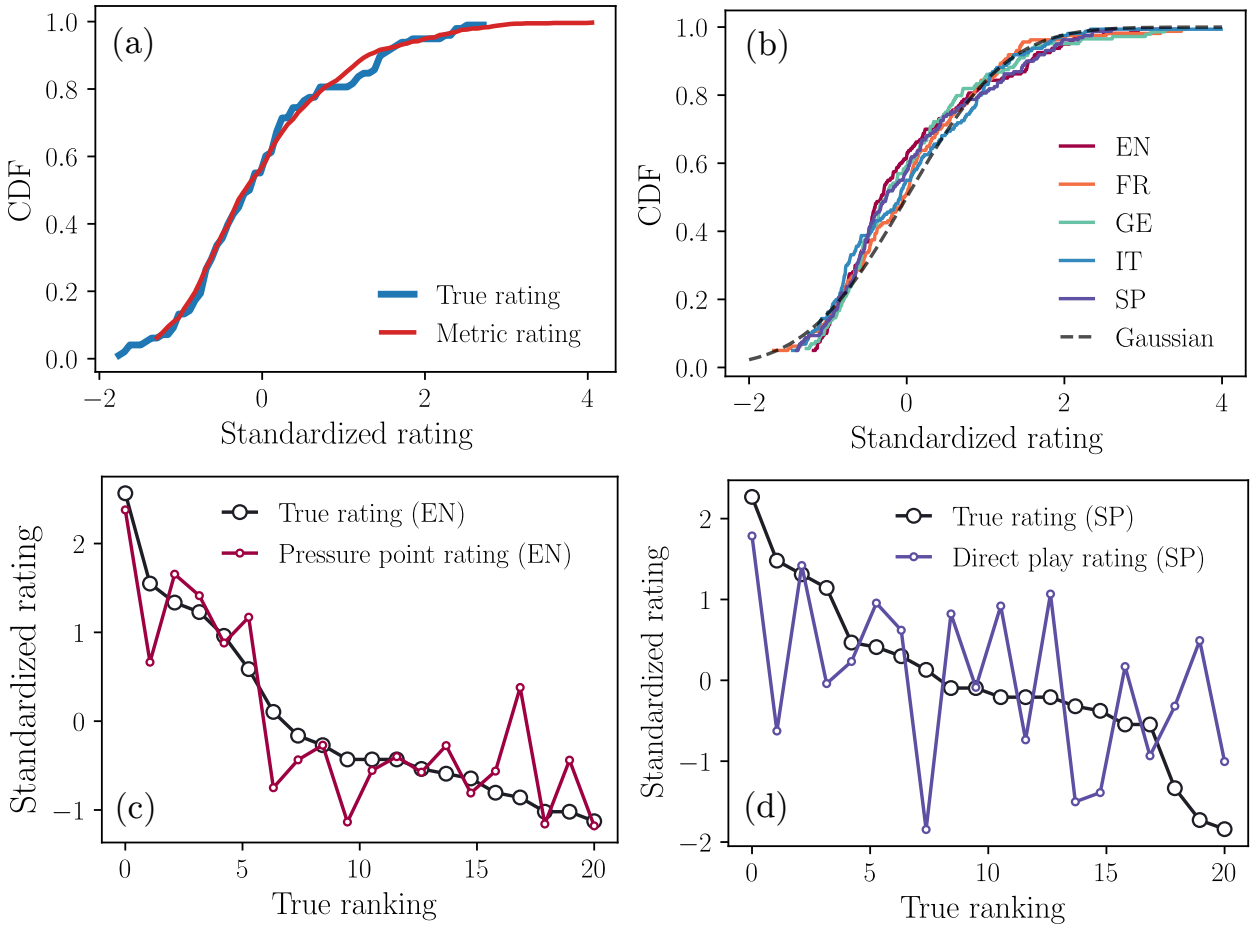

FIG. 2. Comparison between the true and metric-based rating. Since these quantities are expressed in different units, standardized values are used for comparison. (a) Comparison between the cumulative distribution functions associated with the true rating and the metric rating. (b) Cumulative distribution function of the metric rating by league. The black dashed line represents, for reference, the CDF of a standard normal distribution. (c) Comparison for the *Pressure point* metric in the English league. (d) Comparison for the *Direct play* metric in the Spanish league.

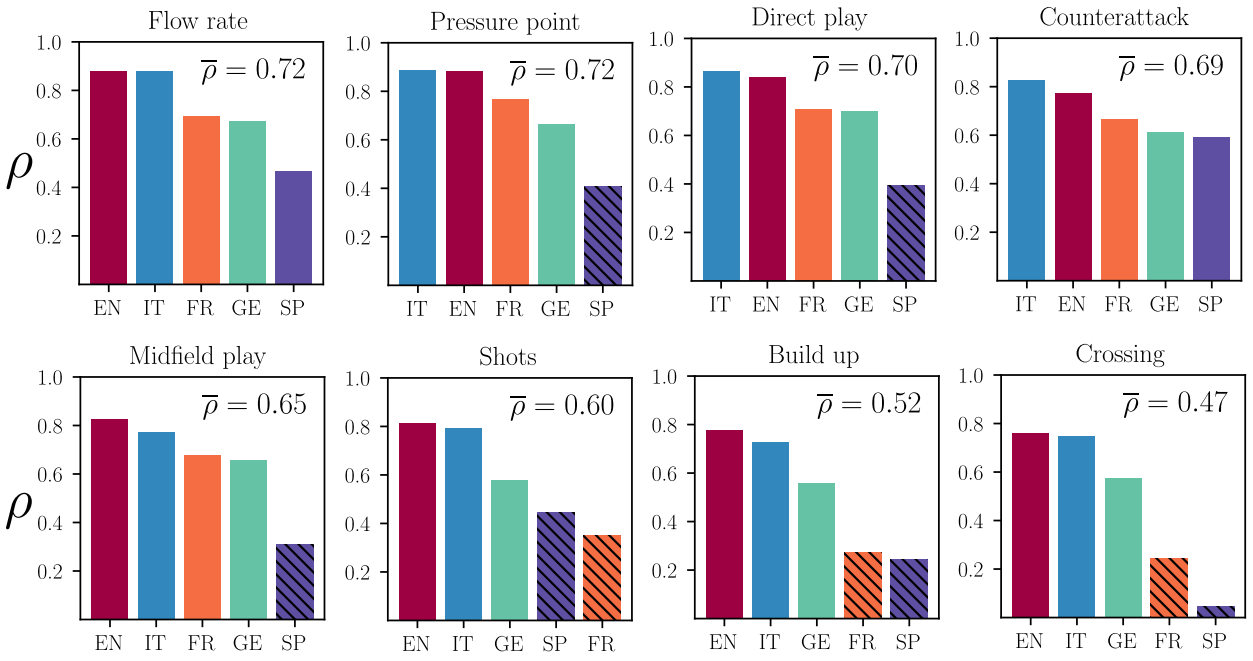

FIG. 3. Pearson correlation coefficient, ρ, between the true rating and the metric-based rating. Each panel corresponds to a performance metric, and the bars show the value of ρ obtained for each league. Within each panel, bars are ordered in increasing order of ρ. Panels are also ordered in increasing order according to the average correlation value for each metric, $\overline{\rho}$. Bars with hatching indicate cases for which the p-value exceeds 0.05. In these cases, no statistically significant correlation was found between the true rating and the metric rating.

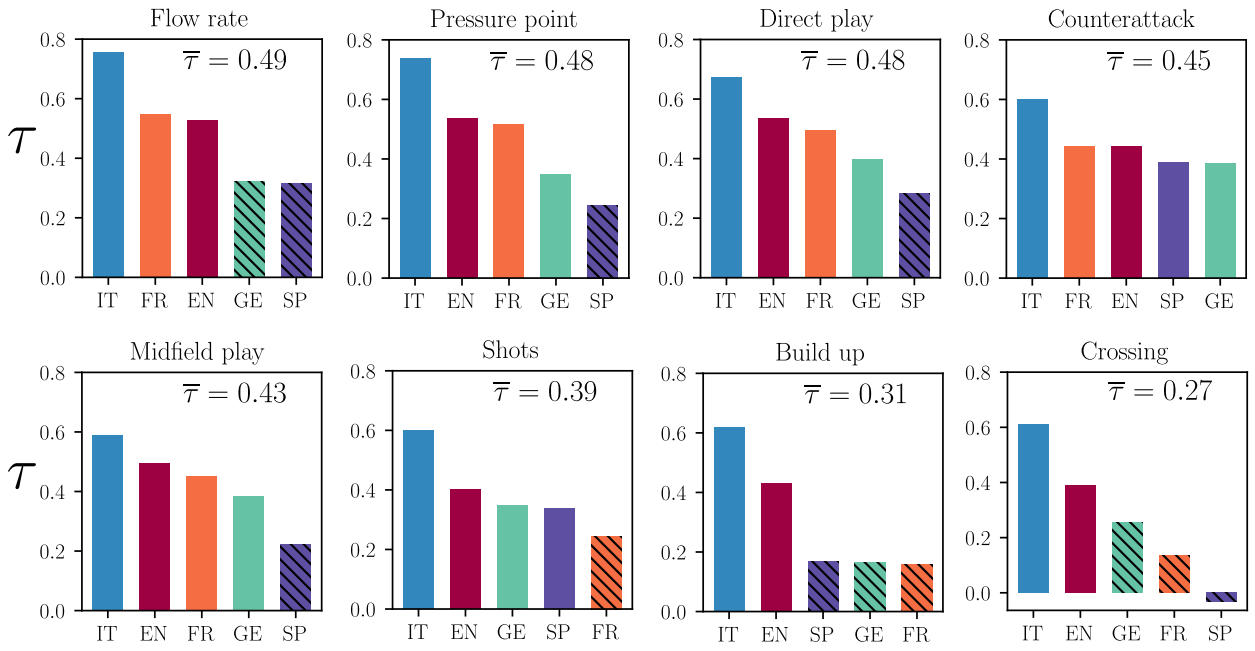

FIG. 4. Kendall rank correlation coefficient, τ, between the true rankings and the rankings derived from performance metrics. Each panel corresponds to a performance metric, and the bars show the value of τ obtained for each league. Within each panel, bars are ordered in decreasing order of τ. Panels are also ordered in decreasing order according to the average Kendall coefficient for each metric, $\overline{\tau}$. Bars with hatching indicate cases for which the p-value exceeds 0.05. In these cases, the results do not provide sufficient evidence to assert the existence of a statistically significant association between the true ranking and the inferred ranking.

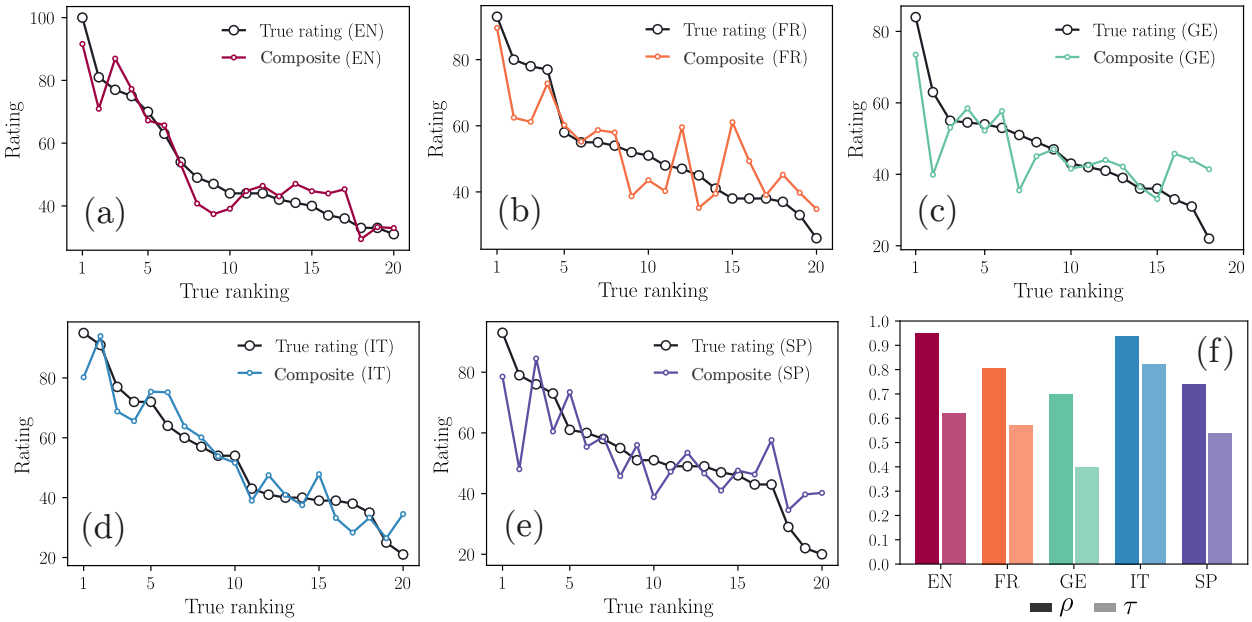

FIG. 5. Comparison between true and composite rating. In panels (a), (b), (c), (d), and (e), we compare the true rating with the composite rating for each team in the English, French, German, Italian, and Spanish leagues, respectively. Panel (f) shows the values of the Pearson correlation coefficient, ρ, and the Kendall rank correlation coefficient, τ, obtained for each league. Note that, in each case, the left bar corresponds to the value of ρ, while the right bar corresponds to the value of τ.

TABLE I. Selected components of the composite rating by league based on BIC.

Metric	**Coef.**	***p*-val**	**Imp. (%)**	R^2
England				
Build up	-24.56	0.0006	37.65	0.88
Flow rate	22.69	0.0045	34.79	
Pressure point	17.98	0.0117	27.57	
France				
Pressure point	33.68	0.0145	61.71	0.61
Direct play	-20.90	0.1096	38.29	
Germany				
Direct play	9.48	0.0012	100	0.46
Italy				
Flow rate	17.84	0.0006	37.15	0.86
Direct play	15.46	0.0018	32.18	
Midfield play	-14.73	0.0119	30.67	
Spain				
Pressure point	-34.93	0.0185	43.08	0.46
Flow rate	26.38	0.0297	32.53	
Counterattack	19.78	0.0046	24.39	