The Journal of Open Source Software

DataBallPy: Load, synchronize, and Analyse your Soccer Data

Gerard Alexander Oonk [1], **Daan Grob**[2], **and Matthias Kempe** [1,3]

1 Department of Sports Sciences, University of Groningen, the Netherlands **2** Independent Researcher, the Netherlands **3** Centre for Sport Science and University Sports, University of Vienna, Austria

DOI: 10.21105/joss.10223

Software

- Review
- Repository
- Archive

Editor: Arfon Smith

Reviewers:

- @felixchenier
- @melund

Submitted: 28 January 2026
Published: 21 April 2026

Summary

Over the last decade, there has been a growing interest in soccer analytics from different backgrounds and for different use cases. Example use cases include: (1) practical decision-making and player benchmarking based on aggregated metrics such as pass success percentage and expected goals (xG) (Goes, Meerhoff, et al., 2020); (2) training periodization and injury prediction using internal and external load metrics (Hader et al., 2019); and (3) behavioral science research focusing on group and subgroup dynamics (Goes, Brink, et al., 2020). The interest in soccer analysis has also increased as data has become more openly available (Bassek et al., 2025). However, a key challenge is that every data provider uses their own data format, which makes it hard to compare and switch between different providers and create large datasets that encompass different leagues and competitions. Currently, open-source packages like Kloppy address this challenge by providing a uniform data format. Similarly, the scientific community has proposed a common data format for soccer game data (Anzer et al., 2025). While Kloppy focuses primarily on parsing soccer data, Floodlight (Raabe et al., 2022) provides a framework for physical analysis of team sports, and mplsoccer is widely utilized for visualizing soccer data.

Recently, there has been a growing interest in combining event and tracking data for contextualized tactical analysis of soccer games. This enables analysts to not only identify when a pass occurred (event data) but also assess the defensive structure during the pass (Forcher et al., 2022; Herold et al., 2022) and evaluate other passing options available at that moment (tracking data) (Spearman et al., 2017). Contextual analysis goes beyond aggregated metrics, enabling quantative analysis of specific moments or phases in the game (Jerome et al., 2024; Oonk, Buurke, et al., 2025). Merging tracking and event data is a key challenge for contextualized analysis of soccer games. `DataBallPy` is an open-source Python package for contextual analysis of soccer games, achieved by: (1) using a standardized data format for both event and tracking data; (2) bundling all game data into a unified framework, rather than treating them as separate objects; (3) incorporating a high-quality, learning-free synchronization algorithm compatible with any combination of tracking and event data providers; and (4) integrating multiple practical and scientific features for efficient computation with minimal user input.

Statement of need

Modern soccer analytics increasingly rely on both event data and tracking data for a more in-depth analysis. Event data captures specific information about events (e.g., passes and shots) including their location, success, start location, and the athlete involved in the action. This information is primarily aggregated for tactical game and player analysis (Goes, Meerhoff, et al., 2020) but is also widely used in scouting because of the low cost and widespread

availability of the data (Arem et al., 2025). In contrast, tracking data provides spatiotemporal information for all athletes and the ball at frequencies ranging between 10 and 25 Hz (Linke et al., 2020). This data is primarily used to quantify physical performance but also for detecting dynamic formations (Sotudeh, 2025), identifying events (Vidal-Codina et al., 2022), classifying game phases (Bauer et al., 2023), analyzing space occupation (Rein et al., 2017; Spearman et al., 2017), and quantifying danger (Link et al., 2016). However, there has been a increasing interest in combining event and tracking data to enrich event information with spatiotemporal context.

This added context provides insights and nuances, primarily on a tactical level, that neither event nor tracking data can provide independently. For example, shot events are enriched with defensive and goalkeeper positioning data to improve expected goals models (Anzer & Bauer, 2021); passes are evaluated using risk-reward assessments of all possible passing options (Goes et al., 2021); determinants of successful 1v1 actions are modeled from spatiotemporal features (Oonk, Buurke, et al., 2025); and the spatiotemporal context of events is used to predict the dangerousity of a game state (Fernández et al., 2021).

A contextual analysis requires a proper synchronization of event and tracking data. Although both event and tracking data provide timestamps, their alignment has been shown to be extremely poor, with reported errors of 1.82 ($\pm$4.06) seconds (Anzer & Bauer, 2021). The random error is particularly concerning because it precludes easy correction; within 4 seconds, the game may have evolved to an entirely different situation, which impacts the stability of the found effects. Oonk, Grob, et al. (2025) demonstrated that the expected goal model decreased in Brier loss from 0.096 to 0.082 (lower is better) when using the synchronized data compared to naive timestamp alignment. Similarly, the feature importance of features that relied on combined tracking and event data information were close to 0 in the naive timestamp synchronization model, unlike the properly synchronized situation (Oonk, Grob, et al., 2025). Thus, there is a need for user-friendly software with a state-of-the-art synchronization algorithm and a convenient data structure for subsequent analysis.

State of the field

Currently available packages enable parsing and analysis of tracking and event data separately. (Kloppy) is a well-known data parsing package in the soccer analytics field. Its primary focus is to simplify and standardize the parsing of soccer tracking and event data from various providers. A similar project aims to establish a standardized format for soccer, which could make different parsers redundant in the future (Anzer et al., 2025). Other packages support analysis (Raabe et al., 2022) and visualization (mplsoccer) of soccer tracking and event data.

Some different open-source projects focus less on parsing and analysis, but do incorporate synchronization between tracking and event data. Roy et al. (2024) synchronizes events with the tracking data using event-specific cost functions. However, the approach takes up to three minutes per match and can leave events unassigned to the tracking data. Kim et al. (2025) recognized that event locations often contain large errors and thus developed an algorithm to merge event and tracking data without relying on event positions. However, both approaches take considerable time, may shuffle events during chaotic situations, and do not provide a convenient data structure for subsequent analysis. `DataBallPy` uses an earlier proposed Needleman-Wunsch algorithm to merge tracking and event data (Kwiakowski & Clark, n.d.) while optimizing it so it runs in mere seconds (Oonk, Grob, et al., 2025) instead of ten minutes. Next to a state-of-the-art synchronization algorithm, `DataBallPy` offers a unified data structure for different data providers and an intuitive data structure for subsequent analysis.

Software Design

A core design choice has been to center DataBallPy around the Game object. Instead of treating data as separate, independent streams (as in existing packages), the Game object consolidates all information for a single game: it consists of metadata, tracking, and event data. This unified approach reinforces the idea that combined information holds more value than independent sources. It also allows for simple usage by creating single methods that rely on all information available on the game without excessive user input. Finally, by combining all information in a single Game object, it becomes more convenient to deliver optimized functionalities. Some of the most important utilities of DataBallPy are listed below.

Parsing Data

DataBallPy parses data from different commercial data providers such as Tracab, Metrica, Inmotio, Opta, Instat, SciSports, Sportec, and Statsbomb internally using the get_game function. The Game object contains the event and tracking data internally as Pandas dataframes, making them intuitive to work with (team, 2020). Alternatively, users can parse data from various providers using Kloppy and convert Kloppy event and tracking datasets into a Game object with the get_game_from_kloppy function. Last, DataBallPy has included a function to load openly available data directly in a Game object using get_open_game, which allows users who do not have access to data to still work with soccer data in DataBallPy (Bassek et al., 2025). As parsing and preprocessing a single game can take between a few seconds and several minutes on a standard device (comparable to other packages), DataBallPy allows users to efficiently save preprocessed Game objects as Parquet or JSON files. This offers two key benefits: (1) preprocessed Game objects can be loaded in milliseconds using get_saved_game, rather than minutes, and (2) raw tracking data files (up to 400 MB per game) are reduced to 20–100 MB when saved as DataBallPy objects, which include both event and tracking data.

Preprocessing

Tracking data typically captured from video footage using computer vision. Depending on the quality and number of cameras, some noise may affect both athlete and ball positions (Linke et al., 2020). DataBallPy allows for filtering of the ball and positional data as well as differentiation of positions to compute velocity and acceleration. Furthermore, tracking data enables computation of individual athlete possession (Vidal-Codina et al., 2022), while combined with event data, team-level possession can be estimated.

synchronization

DataBallPy uses a soccer-specific implementation of the Needleman-Wunsch algorithm to synchronize the event and tracking data, which is more elaborately described in (Oonk, Grob, et al., 2025). The game can be synchronized using the following code

```
>>> from databallpy import get_open_game
>>> game = get_open_game()
>>> game.synchronise_tracking_and_event_data()
```

Performance Indicators

DataBallPy includes an extensive list of scientific features. All features can be computed with minimal code after obtaining a Game object. The documentation provides detailed explanations of the feature-computing code, enabling clear reporting and reproducibility. Using DataBallPy, the following features can be computed:

- Covered Distance (in specific velocity and acceleration zones) (Jerome et al., 2024)
- Pressure (Andrienko et al., 2017; Herold et al., 2022)

- Individual player possession (Vidal-Codina et al., 2022)
- Expected Goals (Anzer & Bauer, 2021)
- Expected Threat (Singh, 2019)
- Voronoi Space Occupation (Rein et al., 2017)
- Pitch Control (Fernandez & Bornn, 2018)
- Dangerous Accessible Space (Bischofberger & Baca, 2025)

Visualisation

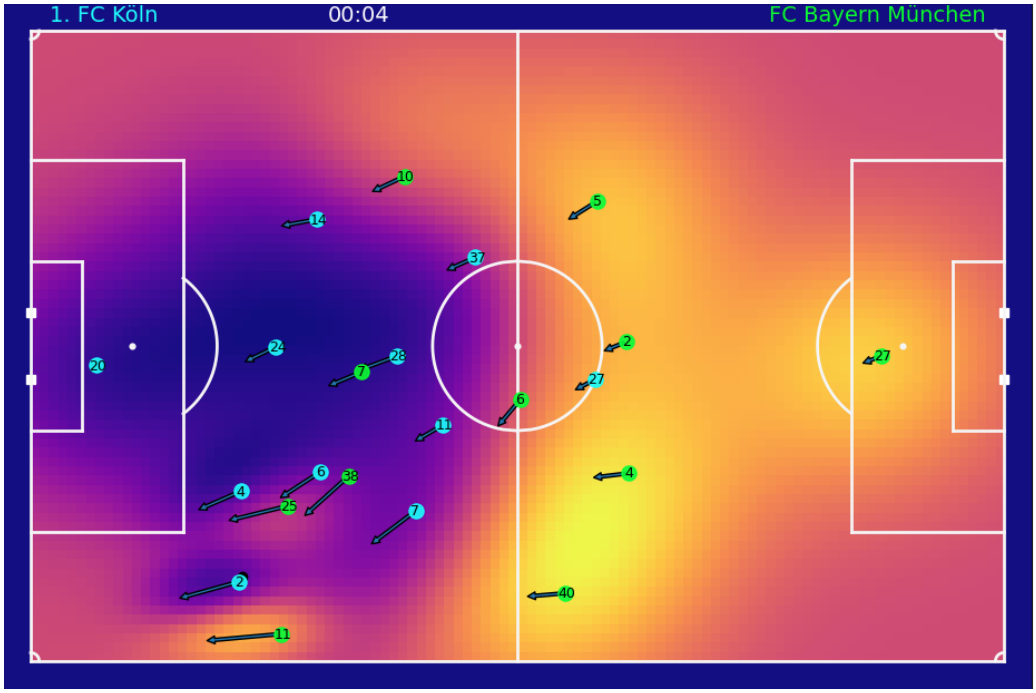

Figure 1: Example plot of soccer tracking data with pitch control heatmap as introduced in Fernandez & Bornn (2018)

`DataBallPy` includes elaborate functionality to visualize the data in the `Game` object. Event locations can be visualized on a pitch using the `plot_events()` function, which allows for coloring of events by outcome, team, or event type during specific periods in the game. Similarly, the locations and velocities of all players can be plotted using the `plot_tracking_data()` function. If the event and tracking data are synchronized, users can overlay event information in the same plot. Other features like pitch control heatmaps, player possession, and any custom feature can also be visualized simultaneously with the event and tracking data (Figure 1). Finally, the tracking data (with heatmaps and custom features) can be transformed into a video (mp4) to show the true spatiotemporal progression over time.

```
import matplotlib.pyplot as plt

from databallpy import get_open_game
from databallpy.visualize import plot_tracking_data

game = get_open_game()
game.tracking_data.add_velocity(game.get_column_ids() + ["ball"])

pitch_control = game.tracking_data.get_pitch_control(
    game.pitch_dimensions,
```

```
    start_idx=100,
    end_idx = 101
)


fig, ax = plot_tracking_data(
    game,
    idx=100,
    add_velocities=True,
    heatmap_overlay=pitch_control[0],
    overlay_cmap="plasma",
    team_colors=["#00FFFF", "#00FF00"]
)
plt.show()
```

Research impact statement

DataBallPy has been increasingly adopted by developers, practitioners, and researchers. The project has over 60 GitHub stars. Issues and PRs are being opened by users outside of the network of the original owners and maintainers. Additionally, DataBallPy has been mentioned in numerous published scientific papers (Anzer et al., 2025; Oonk, Buurke, et al., 2025; Robertson et al., 2023; Zhang et al., 2025). Moreover, the largest currently open-sourced dataset of tracking and event data showcased how DataBallPy can be used to synchronize the two sources (Bassek et al., 2025). Finally, authors who introduce new metrics propose to open a PR with their metric so it is easily available for the scientific community (Bischofberger & Baca, 2025). Collectively, this demonstrates DataBallPy's broad user base and its growth beyond the original mentainers' network.

AI usage disclosure

Generative AI was used for reformulation of sentences in this manuscript. No generative AI tools were used in the development of the core functionalities and architecture of DataBallPy. Except for unittests, there is no explicit restriction on the usage of generative AI in the further development of DataBallPy (e.g., optimizing code, docstrings, reviewing, writing documentation, etc.). All code and documentation are checked and verified by human maintainers before merging into the code base.

References

Andrienko, G., Andrienko, N., Budziak, G., Dykes, J., Fuchs, G., Landesberger, T. von, & Weber, H. (2017). Visual analysis of pressure in football. *Data Mining and Knowledge Discovery*, *31*, 1793–1839. https://doi.org/10.1007/s10618-017-0513-2

Anzer, G., Arnsmeyer, K., Bauer, P., Bekkers, J., Brefeld, U., Davis, J., Evans, N., Kempe, M., Robertson, S. J., Smith, J. W., & Haaren, J. V. (2025). *Common data format (CDF): A standardized format for match-data in football (soccer)*. https://arxiv.org/abs/2505.15820v4

Anzer, G., & Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, *3*, 624475. https://doi.org/10.3389/fspor.2021.624475

Arem, K. van, Goes-Smit, F., & Söhl, J. (2025). Forecasting the future development in quality and value of professional football players. *Applied Sciences*, *15*(16), 8916. https:

//doi.org/10.3390/app15168916

Bassek, M., Rein, R., Weber, H., & Memmert, D. (2025). An integrated dataset of spatiotemporal and event data in elite soccer. *Scientific Data*, *12*, 195. https://doi.org/10.1038/S41597-025-04505-Y

Bauer, P., Anzer, G., & Shaw, L. (2023). Putting team formations in association football into context. *Journal of Sports Analytics*, *9*, 39–59. https://doi.org/10.3233/JSA-220620

Bischofberger, J., & Baca, A. (2025). *Dangerous accessible space: A unified model of space and value in team sports*. https://doi.org/10.21203/RS.3.RS-6932689/V1

Fernandez, J., & Bornn, L. (2018). Wide open spaces: A statistical technique for measuring space creation in professional soccer. *Sloan Sports Analytics Conference*. https://www.researchgate.net/publication/324942294_Wide_Open_Spaces_A_statistical_technique_for_measuring_space_creation_in_professional_soccer

Fernández, J., Bornn, L., & Cervone, D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, *110*, 1389–1427. https://doi.org/10.1007/S10994-021-05989-6

Forcher, L., Forcher, L., Altmann, S., Jekauc, D., & Kempe, M. (2022). The keys of pressing to gain the ball–characteristics of defensive pressure in elite soccer using tracking data. *Science and Medicine in Football,*. https://doi.org/10.1080/24733938.2022.2158213

Goes, F., Brink, M., Elferink-Gemser, M., Kempe, M., & Lemmink, K. A. P. M. (2020). The tactics of successful attacks in professional association football: Large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. *Journal of Sports Sciences*, *39*, 523–532. https://doi.org/10.1080/02640414.2020.1834689

Goes, F., Meerhoff, L. A., Bueno, M. J. O., Rodrigues, D. M., Moura, F. A., Brink, M. S., Elferink-Gemser, M. T., Knobbe, A. J., Cunha, S. A., Torres, R. S., & Lemmink, K. A. P. M. (2020). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, *21*, 481–496. https://doi.org/10.1080/17461391.2020.1747552

Goes, F., Schwarz, E., Elferink-Gemser, M., Lemmink, K., & Brink, M. (2021). A risk-reward assessment of passing decisions: Comparison between positional roles using tracking data from professional men's soccer. *Science and Medicine in Football*, *6*, 372–380. https://doi.org/10.1080/24733938.2021.1944660

Hader, K., Rumpf, M. C., Hertzog, M., Kilduff, L. P., Girard, O., & Silva, J. R. (2019). Monitoring the athlete match response: Can external load variables predict post-match acute and residual fatigue in soccer? A systematic review with meta-analysis. *Sports Medicine - Open*, *5*, 48–48. https://doi.org/10.1186/S40798-019-0219-7

Herold, M., Hecksteden, A., Radke, D., Goes, F., Nopp, S., Meyer, T., & Kempe, M. (2022). Off-ball behavior in association football: A data-driven model to measure changes in individual defensive pressure. *Journal of Sports Sciences*, *40*, 1412–1425. https://doi.org/10.1080/02640414.2022.2081405

Jerome, B. W. C., Stoeckl, M., Mackriell, B., Dawson, C. W., Fong, D. T. P., & Folland, J. P. (2024). Contextualised physical metrics: The physical demands vary with phase of play during elite soccer match play. *European Journal of Sport Science*, *24*, 1627–1638. https://doi.org/10.1002/EJSC.12209

Kim, H., Choi, H., Seo, S., Boomstra, T., Yoon, J., & Park, C. (2025). *ELASTIC: Event-tracking data synchronization in soccer without annotated event locations*. https://arxiv.org/abs/2508.09238

Kwiakowski, M., & Clark, A. (n.d.). *The right way to synchronise event and tracking data*. https://kwiatkowski.io/sync.soccer

Link, D., Lang, S., & Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLOS ONE*, *11*, e0168768. https://doi.org/10.1371/JOURNAL.PONE.0168768

Linke, D., Link, D., & Lames, M. (2020). Football-specific validity of TRACAB's optical video tracking systems. *PLOS ONE*, *15*, e0230179. https://doi.org/10.1371/JOURNAL.PONE.0230179

Oonk, G. A., Buurke, T. J. W., Lemmink, K. A. P. M., & Kempe, M. (2025). The interaction between attacker and environment predicts successfulness in one-on-one dribbles in male elite football. *Journal of Sports Sciences*. https://doi.org/10.1080/02640414.2025.2555117

Oonk, G. A., Grob, D., & Kempe, M. (2025). The right way to synchronize tracking and event data: Using domain knowledge to optimize algorithms. In D. Goossens (Ed.), *MathSports conference* (pp. 136–143).

Raabe, D., Biermann, H., Bassek, M., Wohlan, M., Komitova, R., Rein, R., Groot, T. K., & Memmert, D. (2022). Floodlight - a high-level, data-driven sports analytics framework. *Journal of Open Source Software*, *7*(76), 4588. https://doi.org/10.21105/joss.04588

Rein, R., Raabe, D., & Memmert, D. (2017). "Which pass is better?" Novel approaches to assess passing effectiveness in elite soccer. *Human Movement Science*, *55*, 172–181. https://doi.org/10.1016/j.humov.2017.07.010

Robertson, S., Duthie, G. M., Ball, K., Spencer, B., Serpiello, F. R., Haycraft, J., Evans, N., Billingham, J., & Aughey, R. J. (2023). Challenges and considerations in determining the quality of electronic performance & tracking systems for team sports. *Frontiers in Sports and Active Living*, *5*, 1266522. https://doi.org/10.3389/fspor.2023.1266522

Roy, M. V., Cascioli, L., & Davis, J. (2024). ETSY: A rule-based approach to event and tracking data SYnchronization. *Communications in Computer and Information Science*, *2035 CCIS*, 11–23. https://doi.org/10.1007/978-3-031-53833-9_2

Singh, K. (2019). *Introducing expected threat (xT)*. https://karun.in/blog/expected-threat.html

Sotudeh, H. (2025). The principles of tactical formation identification in association football (soccer)—a survey. *Frontiers in Sports and Active Living*, *6*, 1512386. https://doi.org/10.3389/fspor.2024.1512386

Spearman, W., Basye, A. T., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics-based modeling of pass probabilities in soccer. *Sports Analytics Conference*. https://www.researchgate.net/profile/William-Spearman/publication/315166647_Physics-Based_Modeling_of_Pass_Probabilities_in_Soccer/links/58cbfca2aca272335513b33c/Physics-Based-Modeling-of-Pass-Probabilities-in-Soccer.pdf

team, T. pandas development. (2020). *Pandas-dev/pandas: pandas* (latest). Zenodo. https://doi.org/10.5281/zenodo.3509134

Vidal-Codina, F., Evans, N., Fakir, B. E., & Billingham, J. (2022). Automatic event detection in football using tracking data. *Sports Engineering*, *25*. https://doi.org/10.1007/s12283-022-00381-6

Zhang, G., Kempe, M., McRobert, A., Folgado, H., & Olthof, S. B. H. (2025). Navigating team tactical analysis in football: An analytical pipeline leveraging player tracking technology. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*. https://doi.org/10.1177/17543371251392456