Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

Research on Momentum in Tennis Based on Random Forest

Linjie Tang[1, *], Kangyu Chen[2]

[1]Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China, 518055

[2]Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China, 518055

*Corresponding author: 12111204@mail.sustech.edu.cn

Abstract. In tennis, there is an increasing emphasis on the significance of “momentum,” particularly in comprehending and effectively using this phenomenon in a scientific manner. However, The existence of momentum in competitions is sometimes challenging to scientifically prove. This work strives to develop techniques based on random forest to quantify momentum and predict players’ winning rates. Upon analysis, the paper extracts three features that have a significant impact on the player’s winning rate in a match. Extensive experiments conducted on two real-world datasets from the Wimbledon Championships show the strong performance of our predictive model. Next, other tennis matches like the US Open are selected to make predictions that demonstrate the good ability of generalization. A run test is also conducted to support the model further. Therefore, there is sufficient confidence to assert that the defined momentum is effective. Finally, this work provides a comprehensive overview of the role of “momentum” and offers insightful suggestions for players to enhance their performance on the court.

Keywords: Momentum, Tennis, Random Forest, Run Test.

1. Introduction

In sports, "momentum" refers to a psychological and performance dynamic exhibited by athletes during a competition. This momentum can be triggered by a series of successful actions, scoring, or outstanding performances, leading to elevated levels of confidence and performance. In tennis matches, this phenomenon is more apparent. The leader's momentum is often greater and tends to significantly surpass that of the trailing competitor[1]. Furthermore, numerous studies indicate that there are many critical points in tennis[2]. The performance during these decisive moments, such as break points[3], often causes fluctuations in a player's momentum and frequently determines the outcome of the match.

The concept of momentum is still challenging to observe and define scientifically. There are many definitions, such as dividing momentum into psychological momentum and strategic momentum[4] or defining it as psychological and behavioral momentum[5].

Although there are many qualitative studies on momentum, it has not been quantified and used for prediction. Currently, some have attempted to quantify momentum and make predictions using machine learning, such as ARIMA model refined through logistic regression[6], XGboost[7], and Support Vector Machines[8].

Existing models are typically based on predicting outcomes from best-of-three sets[9] or best-of-five sets[10]. However, in official tennis tournaments, a set usually consists of multiple games, and each game includes several points. The work focuses on predicting these smaller units, points, and games in tennis matches. In the paper, a win probability prediction model based on random forests is developed to quantify strategic momentum. Then, the paper observes the momentum of the players during matches. Finally, the work use a run test to demonstrate that the momentum defined plays a significant role in the match.

2. Random Forest predictive model

2.1. Definition of momentum

In the paper, it is assumed that each player adheres to the rules of fair competition, avoiding sandbagging or any form of cheating while also excluding factors such as venue conditions, weather, players current physical states, referee decisions, and other unexpected events that could affect the states of both players.

Based on experience, if, at a certain moment, a player has greater momentum, then the probability of winning the next point or game is higher. Thus, their performance is better. Momentum is also considered to be similar to a pendulum. When the momentum of one side is high, the other side's momentum is low, and the speed of momentum rising and falling is rapid. The definition of 'Momentum' is as follows:

$$V_i^* = \theta_1 Vg_i + \theta_2 Vp_i \tag{1}$$

$$M = V_1^* - V_2^* \tag{2}$$

where Vg_i is the winning rate of player i for the current game, Vp_i is the winning rate of player i for the current point, θ_1, θ_2 are the weight of Vg_i and Vp_i respectively, and M is momentum. It should be noted that when player 1 gets the advantage, he or she gains a momentum closer to 1. When player 1 gets a disadvantage, the momentum would be closer to -1.

2.2. Analysis of features influencing winning rates

Data analysis is performed on four players' attributes and the game's winning rate. The three attributes include the scoring pattern in the first three points of one game, the scoring pattern in the preceding three points before one point, and the server of one point. A detailed explanation of the features that impact momentum will be provided next.

(1) Scoring pattern in the first three points of one game

Firstly, it is believed that the first three points of a match may influence a player’s momentum in that game. In the early stages, the leading player often gains a psychological advantage, especially when leading in the first three points. This lead can boost a player’s confidence and enhance their performance. The paper defines the result of a point as 1 for Player 1 winning and 2 for Player 2 winning. The outcomes of the first three points can be represented by a tuple containing three elements, denoted by ”pat.” There are eight possible patterns, as detailed in Table 1.

Table.1. Pattern tuple

Pat	Tuple	Pat	Tuple	Pat	Tuple	Pat	Tuple
1	(1,1,1)	3	(1,1,2)	5	(2,1,1)	7	(2,1,2)
2	(2,2,2)	4	(1,2,1)	6	(1,2,2)	8	(2,2,1)

The pattern-winning rate graph for player 1 and player 2 is shown in fig.1.

Observations show that any player's probability of winning the game after capturing the initial three points exceeds 95%. Assuming both players have an equal winning rate of 50%, the calculated winning rate is around 90%. Conversely, the probability of losing three consecutive points and then making a comeback is about 3%. These statistics indicate that the outcome of the first three points can significantly influence the winning rate, enhancing the momentum of the leading player while diminishing it for the trailing player.

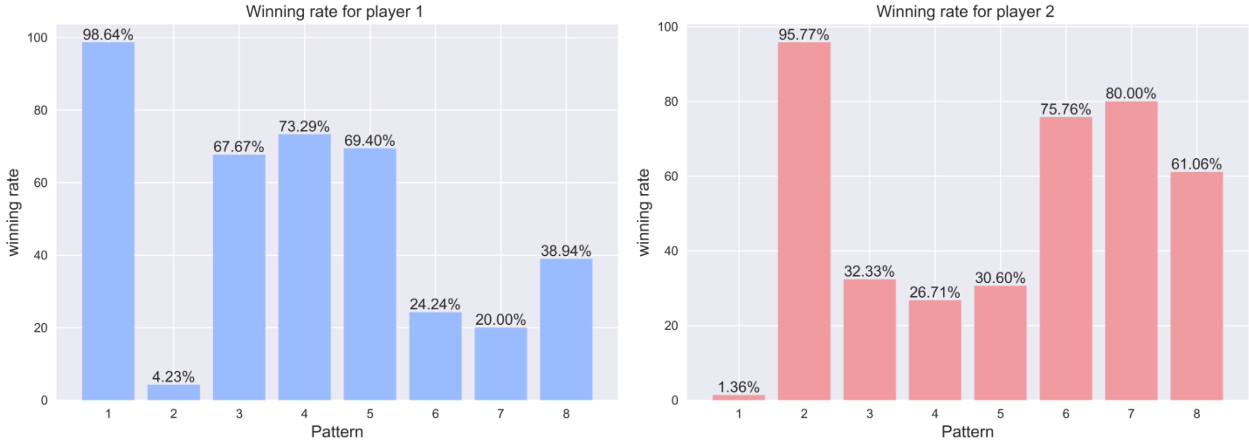

Fig.1 Pattern and winning rate of play 1 and player 2

(2) Scoring pattern in the preceding three points before one point

Similar to (1), this paper believes that the winning probability of the current point or game is influenced by the scores of the preceding points. If a player wins in the previous few points, it may boost his confidence and momentum. He will perform better in the following points.

Since the number of points in a tennis game is relatively small, and it is assumed that players are only influenced by short-term outcomes, the scoring pattern in the preceding three points before one point is defined as the "pre-pattern," denoted as "ppat". For the first two points of each game, as it is the beginning of a new game, the player's mental and physical condition may be different from the previous game. There is also some greater randomness. Therefore, the pre-pattern of this situation is recorded as 0. For the third point of each game, ppat is defined as the score of the two points before this point. The specific formula for the pre-pattern is shown in Table 2.

Table.2. Pre-pattern tuple

Ppat	Tuple	Ppat	Tuple	Ppat	Tuple	Ppat	Tuple
-4	(2,2)	-1	(1,1)	3	(1,1,2)	6	(1,2,2)
-3	(2,1)	1	(1,1,1)	4	(1,2,1)	7	(2,1,2)
-2	(1,2)	2	(2,2,2)	5	(2,1,1)	8	(2,2,1)
0	The point is either the first or second point of the game						

The pre-pattern and winning rate graph for players is shown in fig.2. Observations indicate that, regardless of the player, winning two or more games in the previous three or two points leads to a high winning rate for that game. This shows that prior success effectively improves a player's momentum, leading to their victory in the game.

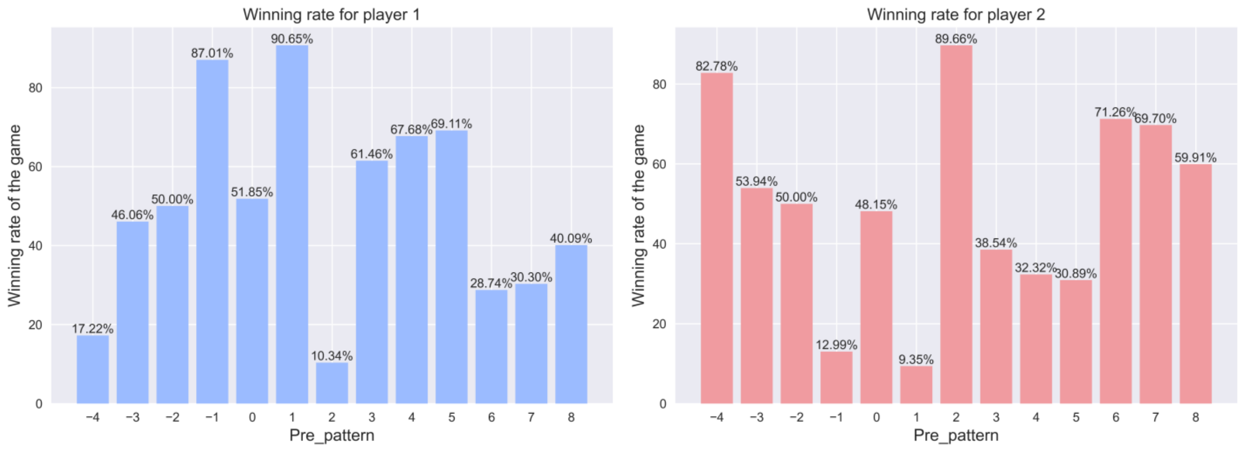

Fig.2 Pre-pattern and winning rate of play 1 and player 2

(3) Server of the point

In tennis, serving is often an important means for players to get a score. According to the tennis rules, the same person serves throughout a game, so the player serving has a significantly higher chance of winning the game. The server is typically in an attacking position during serves, and this proactive stance can help them gain momentum. The opponent may feel passive and pressured, leading to a decrease in momentum and making errors during returns.

The relationship between the winning rate of player 1 and whether they serve in that game is plotted in fig.3. It is evident that the serving player has a significantly higher probability of winning the game compared to the receiving player, with a nearly 70% difference in winning rates. Therefore, whether a player serves or not also has a significant impact on measuring momentum.

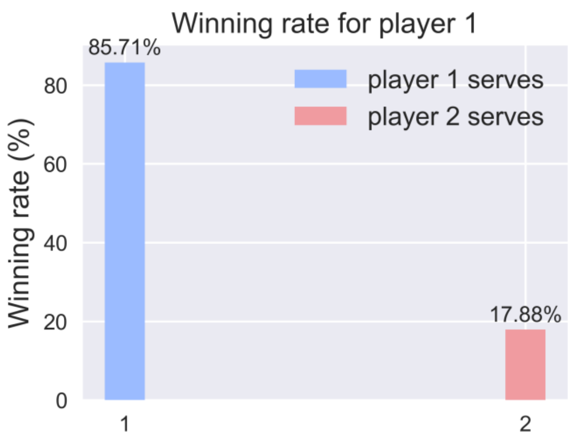

Fig. 3 Server and winning rate of play 1 and player 2

2.3. Data Description

The player ranking data is sourced from ATP and WTA. The ATP, which stands for "Association of Tennis Professionals," is primarily responsible for the management and organization of men's professional tennis. Similarly, the WTA, short for "Women's Tennis Association," is an international organization dedicated to the management and promotion of women's professional tennis.

The training data is sourced from The Match Charting Project (MCP), a volunteer-driven initiative aimed at collecting detailed data on professional tennis matches. The project includes data from both men's and women's matches across various surfaces and different types of tournaments. Each match provides point-by-point statistical information, including shot types, shot directions, return depths, error types, and more.

2.4. The establishment of Random Forest model

Let $x_i = (pat, ppat, sv)$ when training the model, 1323 tennis matches from the Noumea Challenger and US Open datasets are used as the training set, and the 31 matches from Wimbledon are used as the test set. Next, two Random Forest models are employed to predict variables $\mathrm{Vp_i}$ and $\mathrm{Vg_i}$, followed by the calculation of V_i^*. The momentum M is then computed. For data visualization, two matches are selected.

3. Results

3.1. The visualization of two competitions

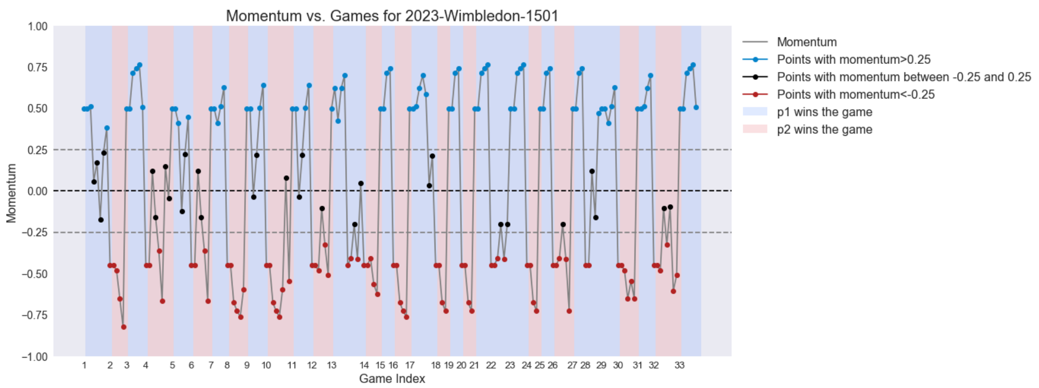

Fig. 4 Visualization of 2023-wimbledon-1501

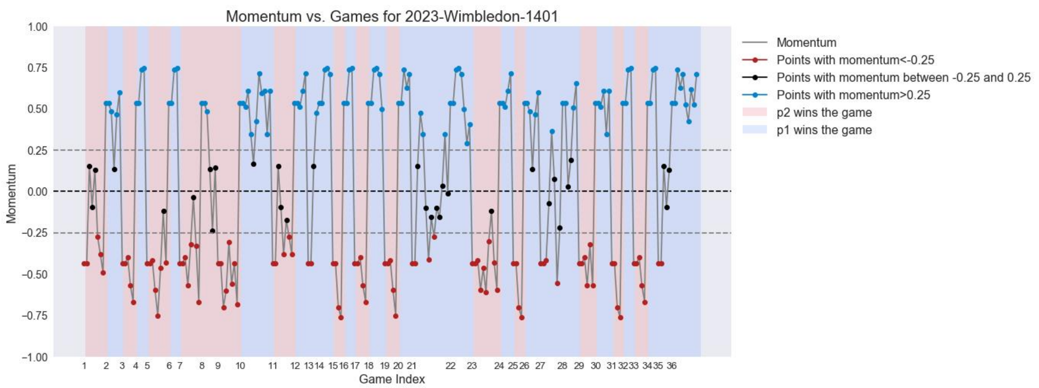

Fig. 5 Visualization of 2023-wimbledon-1401

The fig.4 and fig.5 show how momentum changes as the game progresses in the match '2023-wimbledon-1501' and '2023-wimbledon-1401'. It can be seen that in most cases, at a certain point, when momentum is closer to 1, player 1 is more likely to win; when momentum is closer to -1, player 2 is more likely to win; and when momentum is in [-0.25,0.25], the prediction effect on the game result is not obvious. Meanwhile, the fluctuation of Momentum and the fluctuation of the game winner show basic consistency, regularly fluctuating as the game progresses. All these findings can point out that momentum can be a good portrayal of whether a player of one side is right in a favorable situation in the game, and also explains the alternation of game victors in the game.

3.2. Analysis of experimental results

Given that the prediction of winning rates constitutes a binary classification task, four metrics—accuracy, precision, recall, and F1-score—are employed to assess the efficacy of the model. The performance metrics are delineated in the following Table 3:

Table.3. Performance of the model on the Wimbledon dataset

Predict the victor	Accuracy	Precision (p_1)	Recall (p_1)	F1-Score
A game	81.82%	0.83	0.81	0.82
A point	64.15%	0.69	0.66	0.67

For accuracy, the predicted accuracy is above 60%, indicating that the model can accurately predict the player's winning rate and momentum based on player attributes. This demonstrates the effectiveness of the Random Forest algorithm in this problem.

Overall, the model performs better when predicting the winner of an entire match, exhibiting higher accuracy, precision, recall, and F1-score. In contrast, its performance is poorer when predicting the winner of individual points, with all metrics showing a decline. This may be due to the greater impact of uncontrollable factors in predicting individual points, such as the player's immediate state and minor tactical changes, which are averaged out or offset in predictions throughout an entire match.

3.3. Model generalization capabilities

During the training process, only data from Wimbledon matches, which are played on grass courts, unlike other tournaments conducted on clay or hard courts, is used to train the model. This may affect the style and outcome of the matches. Utilizing Wimbledon data could lead to a model biased towards this specific format of the game. Therefore, it is necessary to test the model's generalization capabilities further, and there is hope to extend this model to other types of ball games in future research.

Two additional competitions have been selected for evaluation: the 2023 US Open (containing 93 matches) and the 2020 Noumea Challenger (containing 119 matches). The model's ability to accurately predict the game victor in these competitions will be assessed, as demonstrated in Tables 4 and 5:

Table.4. Performance of the model on the US Open dataset

Accuracy	Precision (p_1)	Recall (p_1)	F1-Score
81.82%	0.83	0.81	0.82

Table.5. Performance of the model on the Noumea Challenger dataset

Accuracy	Precision (p_1)	Recall (p_1)	F1-Score
81.82%	0.83	0.81	0.82

4. Run test for the model

4.1. The establishment of run test

The run test is a statistical method used to examine whether a data sequence has a trend or pattern. It can be employed to assess whether a sequence demonstrates randomness. Its main idea is to decompose a data sequence into some consecutive runs. A run is a series of consecutive increasing or decreasing values(e.g., 0 or 1). The length of a run is determined by the number of consecutive increasing or descending values.

The test statistic is:

$$Z = \frac{R - \overline{R}}{s_R} \tag{3}$$

where R is the number of observed number of runs, $\overline{R}$ is the expected number of runs, and s_R is the standard deviation of the number of runs. The formula of $\overline{R}$ and s_R are computed as follows:

$$\overline{R} = \frac{2n_0n_1}{n} + 1 \tag{4}$$

$$s_R = \sqrt{\frac{2n_0n_1(2n_0n_1 - n)}{n^2(n-1)}} \tag{5}$$

where n_1 denotes the number of ”1” in the series, n_0 denotes the number of ”0” in the series, and $n = n_0 + n_1$.

In the previous section, momentum is defined. The original score and momentum data sequence are then encoded into 0s and 1s to generate a new sequence. It is defined that when player 0’s momentum exceeds player 1’s and a point is scored, this situation is encoded as "1". When player 0’s momentum exceeds player 1’s but no point is scored, this situation is encoded as "0".

The null hypothesis H0 is that swings in play and runs of success by one player are random. The alternative hypothesis H1 is that swings in play and runs of success by one player are not random.

When a sequence is indeed random, 0 and 1 will present a uniform distribution in the sequence. In other words, even if the player’s momentum is significant, it does not necessarily guarantee an improvement in the player’s performance. If a sequence has a trend, the number of runs will be small; for example, when momentum can improve a player's performance, the number of 1s will be much larger than that of 0s. If a sequence has some periodicity, there will be many runs, for example, 1010101010. Any discernible trend or periodicity suggests a lack of randomness. Therefore, the number of runs in a random sequence should be greater than that in a trend sequence but less than that in a periodic sequence.

4.2. Results of run test

Considering that both n_0 and n_1 are greater than 10 in the problem, the test statistic will be approximately compared to a standard normal distribution table. In other words, under the 0.05 significance level, a test statistic with an absolute value exceeding 1.96 suggests there are grounds to reject the null hypothesis H_0.

For all the matches collected, a run test is conducted. The calculated test statistics are displayed in Table 6. The statistic is much larger than 1.96, so the null hypothesis is rejected at the significance level 0.05. This indicates that momentum plays a significant role in the tennis match.

Table.6. Performance of the run test

$\lvert Z\rvert$	$\lvert Z_{0.95}\rvert$	$\lvert Z_{0.975}\rvert$	$\lvert Z_{0.995}\rvert$
50.127	1.645	1.96	2.576

5. Some inspiration for players

If coaches could predict the future direction of momentum in advance, they would be able to adjust their players' strategies accordingly. This research aims to do precisely that. By collecting a subset of points during a match, it is possible to analyze the current momentum and make relatively accurate predictions about future momentum trends. In terms of predictions for momentum, there are some suggestions for coaches and their players.

Firstly, based on the trend chart, predictions can be made that help players mentally prepare for future changes in win rates. Strengthening training and adjusting mindset during periods of weakness will enable players to better cope with the challenges in the next game. This approach facilitates psychological preparedness and training readiness.

Simultaneously, attention should be given to those points where the player's win rate exhibits periodic declines. Often, at these nodes, the player's performance may be suboptimal, or they may lack initiative, being entirely dictated by the opponent's rhythm. In such situations, players should strategically consider taking breaks during the game, such as requesting a change of balls or clearing the court, to adjust their own state and rhythm.

Last but not least, through the trend chart, players are informed that the situation on the field is ever-changing. They need to be psychologically prepared for underestimation, being out of form, and disruptions in rhythm, along with corresponding tactical responses. Simultaneously, attention should be directed towards intervals where their win rates consistently rise. Seizing opportunities in these intervals allows players to control the rhythm of the game, maintaining a better-paced trajectory towards victory.

6. Conclusions

In conclusion, this research has successfully quantified momentum in tennis by using a random forest model and predicted its impact on match outcomes. The findings demonstrate that momentum is not random and plays a crucial role in the dynamics of a tennis match, influencing players' performance and match results. This insight provides valuable strategies for coaches and players, emphasizing the importance of managing momentum shifts during matches to gain a competitive edge. The success of this model also highlights the potential of machine learning techniques in sports analytics to provide deeper insights into game behavior. In future studies, more player characteristics can be used for training, such as the direction of the serve, whether they break the serve, and running distance. Furthermore, whether this model can be applied to other sports domains requires further exploration.

References

[1] Page L. The momentum effect in competitions: field evidence from tennis matches[C]//Econometric Society Australasian Meeting. 2009.

[2] Meffert D, Breuer J, Ohlendorf L, et al. Towards an understanding of big points in tennis: Perspectives of coaches, professional players, and junior players[J]. Journal of Physical Education and Sport (JPES), 2021, 21(2): 728-735.

[3] Seidl R, Lucey P. Live Counter-Factual Analysis in Women's Tennis using Automatic Key-Moment Detection[C]//Proceedings of the MIT Sloan Sports Analytics Conference. Boston, USA: 2022.

[4] Meier P, Flepp R, Ruedisser M, et al. Separating psychological momentum from strategic momentum: Evidence from men's professional tennis[J]. Journal of economic psychology, 2020, 78: 102269.

[5] Hubbard T L. The varieties of momentum-like experience[J]. Psychological Bulletin, 2015, 141(6): 1081.

[6] Meng X, Tian L, Zhang L. Unveiling Momentum Dynamics in Tennis[J]. Advances in Engineering Technology Research, 2024, 10(1): 586-586.

[7] Lin J, Shao P, Zhang Q. Advancing Tennis Analytics: Comprehensive Modeling for Momentum Identification and Strategic Insights[J]. International Journal of Computer Science and Information Technology, 2024, 2(1): 104-117.

[8] Zhang J, Tang K, Zhu X. Research On the Role and Influencing Factors of Momentum in Tennis Matches Based on Linear Conditional Probability Model and Support Vector Machine[J]. Highlights in Science, Engineering and Technology, 2024, 93: 325-333.

[9] Depken C A, Gandar J M, Shapiro D A. Set-level strategic and psychological momentum in best-of-three-set professional tennis matches[J]. Journal of Sports Economics, 2022, 23(5): 598-623.

[10] Depken C A, Gandar J M, Shapiro D A. Set-level strategic and psychological momentum in best-of-five matches in professional tennis[J]. Applied Economics Letters, 2023, 30(6): 735-739.