Mathematical and Computational Applications

MDPI

Article

A Study of Tennis Tournaments by Means of an Agent-Based Model Calibrated with a Genetic Algorithm

Salvatore Prestipino [1] and Andrea Rapisarda [1,2,3,*]

[1] Dipartimento di Fisica e Astronomia "Ettore Majorana", Universitá di Catania, 95100 Catania, Italy; salv.prestipino@gmail.com
[2] INFN Sezione di Catania, 95100 Catania, Italy
[3] Complexity Science Hub, 1080 Vienna, Austria
* Correspondence: andrea.rapisarda@ct.infn.it

Abstract: In this work, we study the sport of tennis, with the aim of understanding competitions and the associated quantities that determine their outcome. We construct an agent-based model that is able to produce data analogous to real data taken from Association of Tennis Professionals (ATP) tournaments. This model depends on three parameters: the talent weight, the talent distribution width, and the chance distribution width. Unlike other similar works, we do not fix the values of these parameters and we calibrate the model results with the help of a genetic algorithm, thus exploring all possible combinations of parameters in the parameter space that are able to reproduce real system data. We show that the model fits the real data well only for limited regions of the parameter space. Limiting the region of interest in the parameter space allows us to perform further calibrations of the model that give us more information about the competition under study. Finally, we are able to provide useful information about tennis competitions, obtaining quantitative information about all of the important parameters and quantities related to these competitions with very limited a priori constraints. Through our approach, differing from those of other works, we confirm the importance of chance in the studied competitions, which has a weight of around 80% in determining the outcome of tennis competitions.

Keywords: tennis data; agent-based model; genetic algorithm

Citation: Prestipino, S.; Rapisarda, A. A Study of Tennis Tournaments by Means of an Agent-Based Model Calibrated with a Genetic Algorithm. *Math. Comput. Appl.* **2024**, *29*, 77. https://doi.org/10.3390/mca29050077

Academic Editors: Alexandre Souto Martinez and Oliver Schütze

Received: 12 June 2024
Revised: 7 September 2024
Accepted: 9 September 2024
Published: 11 September 2024

1. Introduction

The study of competitions is an important topic concerning different scientific fields, including physics [1,2], biology [3], economics [4], etc. This interest is justified by the fact that competition is a common phenomenon in nature and society, and so it is important to understand its mechanics. When we think about the outcome of competitions, we assume that they are determined almost exclusively by the talents, properties, or inclinations of the competitors, but this is not the case, as general studies have recently demonstrated [5,6]. A lot of other work has been carried out to understand the recipes for success [7–10], and it seems that the outcome of a competition depends on a series of factors, among which the action of chance, represented by all the events that cannot be predicted a priori, plays a major role.

In order to better understand competitions, we want to focus on sports, which gives us access to a large amount of data. For this type of competition, previous studies [11–15], despite using different approaches, have again shown that the talent and preparation of athletes are not the only factors determining the outcome of competitions, and that the action of chance is an important component of competitions.

To study the phenomenon of sports competitions, previous works [11–13] have developed agent-based models [16], a type of model used to study complex systems [17,18] in a wide range of scientific fields [19–21]. These models are useful for reproducing the

relationships between simulated agents in a virtual environment composed of a set of rules. In these works, the calibration of these models was obtained by fixing some parameters with reasonable values extrapolated from previous works, and with these constraints, the models were able to provide information regarding the role of chance in opposition to that of talent in determining the outcome of competitions.

In this paper, we also want to construct an agent-based model; however, we want to avoid constraining the parameters of the model in order to obtain a broader view of the studied competitive phenomena, obtaining information on all the parameters that are important for the competition, and then comparing these results with those used in previous studies.

In particular, in this paper, we focus on the sport of tennis using data from Association of Tennis Professionals (ATP) tournaments [22–24]. Tennis is a sport in which two individuals compete against each other, thus representing an example of direct competition, where the two competitors are somehow linked to each other. In fact, in this type of competition, an event that favors one competitor will at the same time disadvantage the opposing competitor.

The agent-based model for a direct 1 vs. 1 competition, constructed in this paper for the simulation of tennis matches and tournament data, depends on three parameters, namely the weight of talent, the standard deviation of the talent distribution, and the standard deviation of the chance distribution. Thus, we infer the shape of the talent and chance distributions, which are chosen to be Gaussian and symmetric with respect to zero, but we do not impose any other constraints on the possible values of the parameters.

In fact, we use a genetic algorithm [25], a relatively simple yet powerful evolutionary algorithm that can be adapted to be used in conjunction with an agent-based model. By using the genetic algorithm and related computations, we explore a large number of possible parameter combinations to calibrate the agent-based model based on the real data.

Then, in order to better understand the results of our calibration of the parameters, we fix the talent distribution used in previous works to reject some unrealistic results and thus obtain information about the role of chance and talent encoded in the values of the parameters used in our model.

2. Materials and Methods

2.1. Data Preparation

In order to construct a model of the direct competition in the sport of tennis, we needed to prepare the data in a way that highlighted the properties of the system. In particular, the data used in this paper consisted of 108,411 matches played in 2125 tournaments, comprising men's singles matches in ATP tournaments from 1991 to 2021, including the Grand Slams, Masters 1000, ATP 500, and ATP 250. These data were acquired using online resources [22–24].

We sought to estimate the performance of the two competitors involved in a match, and with this information it was then possible to aggregate the performance of all players involved in all matches to construct a distribution of the performance that we could then try to reproduce with a model.

To understand how this performance value is constructed in the case of tennis matches, it is necessary to briefly summarize how points are awarded in this sport [26,27].

- In tennis, the winner of a match is the player who wins at least two of three sets in a three-set match, or wins at least three sets in a five-set match.
- A set is won by winning at least six games and leading the opponent by at least two points. If both players win at least six games, thus ending in a tie, a tie-break is played, which consists of an extra game to decide the set winner.
- A game is won by the first player to score 4 points, with an advantage of at least 2 points over their opponent; if this is not the case, the game continues until one player scores 2 points more than their opponent and wins the game.

The data at our disposal provide us with a lot of statistical information regarding a match: the winning and losing players, the total score, the scores of the winning and losing players, the tournament in which the match was played, the phase of the tournament in which the match was played, the number of games won by the winner and the loser, the number of first serves won by the winner and the loser, etc.

However, we constructed our numerical estimate of performance using a very small subset of this information, In particular, we used the total number of points scored during the match by the two competitors involved, as well as the points scored individually by the winning and losing players. In fact, the score can be seen as the result of the total balance of "positive" and "negative" actions performed by the two competitors during the match, as all the information about their performance is contained in it.

Therefore, we used the total score, the score of the winning player, and the score of the losing player to calculate a normalized performance value (expressed by a number between 0 and 1). We used the ratios between the scores of the two competitors and the total number of points scored by both during the match and obtained two performance values, one for the winning player and one for the losing player. By defining the performance in this way, the value obtained for the winner of the match was linked to the value obtained for the losing player in a specular manner, so if the performance of one player was underwhelming, the performance of the other was enhanced, meaning that decoupling the two performances would be extremely difficult.

All of these performance values constructed for the matches can be accumulated in a distribution of the score performance, as shown in Figure 1a. In this figure, we see two histograms, with one representing the winning players and the other one representing the losing ones. The symmetry of these distributions with respect to 0.5 is significant, as the distance and the intersection of the two distributions in relation to each other are features that we aimed to reproduce with the agent-based model.

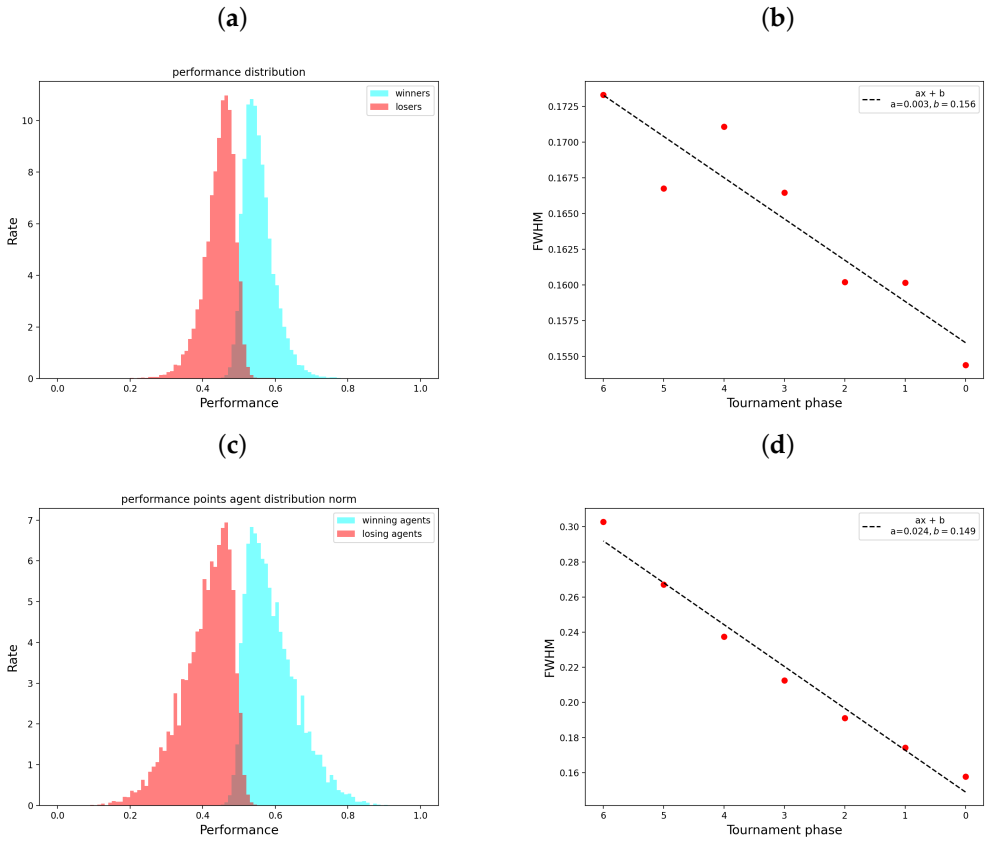

Figure 1. Characteristics of the real and simulated data, highlighting the similarity of the data obtained. (**a**) Histogram constructed by taking into account the score performance values obtained

from the 108,411 matches played in ATP tournaments from 1991 to 2021. The values for the winning players are shown in blue, while the values for the losing players are shown in red. (**b**) Trend of the FWHM of the score performance distribution with respect to the stages of the tournaments. A number is assigned to each stage, with 0 being assigned to the final, 1 to the quarter-finals, etc. A linear fit is also shown to highlight the trend of the FWHMs. (**c**) Score performance distribution obtained using the agent-based model to simulate 10,000 tournaments with 128 agents participating in each, with fixed parameters equal to $a = 0.3$, $\sigma_t = 0.2$, $\sigma_c = 0.2$. (**d**) Trend of the FWHMs of the score performance distributions for the different stages of the tournaments for the same number of simulated tournaments and parameters.

Another feature of the data was highlighted when we studied how the width of the score performance distribution changes across the different stages of the tournaments. The width of the distribution contains information concerning the difference between the score performance of the winning players and that of the losing players. We calculated the full width at half maximum (FWHM) considering the normalized histograms of the score performances of the winning and losing players and performed a Gaussian fit on them.

Figure 1b shows that from the first round, labeled with $x = 6$, to the final, labeled with $x = 0$, the FWHM of the score performance distribution tends to decrease with a certain slope, representing another feature that we want to reproduce with the agent-based model.

2.2. The Direct Competition Model Core

In order to develop an agent-based model for the sport of tennis, it is necessary to establish the mechanism that allows an agent-player to score a point. This mechanism must depend on the sporting performance of the agent-player in such a way to emulate a real situation. Previous works [9,11,12] have used an equation to obtain a numerical value of the sporting performance of an agent-player. We adapted this equation to the case of a direct competition between two agent-players and obtained the following system of equations:

$$\begin{aligned} \widetilde{P}_1^{[a,\sigma_t,\sigma_c]}(t) &= \frac{1}{2} + \frac{d_{1,2}^{[a,\sigma_t,\sigma_c]}(t)}{2} \\ \widetilde{P}_2^{[a,\sigma_t,\sigma_c]}(t) &= \frac{1}{2} - \frac{d_{1,2}^{[a,\sigma_t,\sigma_c]}(t)}{2} \end{aligned} \tag{1}$$

with

$$d_{1,2}^{[a,\sigma_t,\sigma_c]}(t) = \quad a(T_1^{[\sigma_t]} - T_2^{[\sigma_t]}) + (1-a)[2C(t)^{[\sigma_c]} - 1] \tag{2}$$

where $d_{1,2}^{[a,\sigma_t,\sigma_c]}(t)$ is the distance in terms of performance between the two competitors, namely the agent with the label 1 and that with the label 2. This quantity depends on the variable t, the iteration for which the quantity is calculated, and on a series of parameters, as follows:

- The talent weight a, defined with values between 0 and 1, which determines the importance of talent over chance; for $a = 1$, performance is purely talent-dependent, while for $a = 0$, performance is purely chance-dependent.
- The standard deviation of the talent distribution σ_t, with a normal distribution, centered at 0.5 and truncated between 0 and 1, from which the constant values T_1 and T_2, the talent values of agents 1 and 2 involved in a match, are drawn.
- The standard deviation of the chance distribution σ_c, with a normal distribution, centered at 0.5 and truncated between 0 and 1, from which the chance value $C(t)$, recalculated at each iteration, is drawn.

Using $d_{1,2}^{[a,\sigma_t,\sigma_c]}(t)$, it is possible to obtain the values of $\widetilde{P}_1^{[a,\sigma_t,\sigma_c]}(t)$ and $\widetilde{P}_2^{[a,\sigma_t,\sigma_c]}(t)$, which are the relative sporting performance of agents 1 and 2, respectively. Figure 2 shows an

example of the types of values generated by $d_{1,2}^{[a,\sigma_t,\sigma_c]}(t)$; this function, at each iteration, returns a value around $a(T_1^{[\sigma_t]} - T_2^{[\sigma_t]})$, which is the center of the distribution $d_{1,2}^{[a,\sigma_t,\sigma_c]}(t)$. The location of the center depends on the difference in talent between the two agents considered, and the amplitude of this distribution depends on a and on the chance distribution $C(t)$.

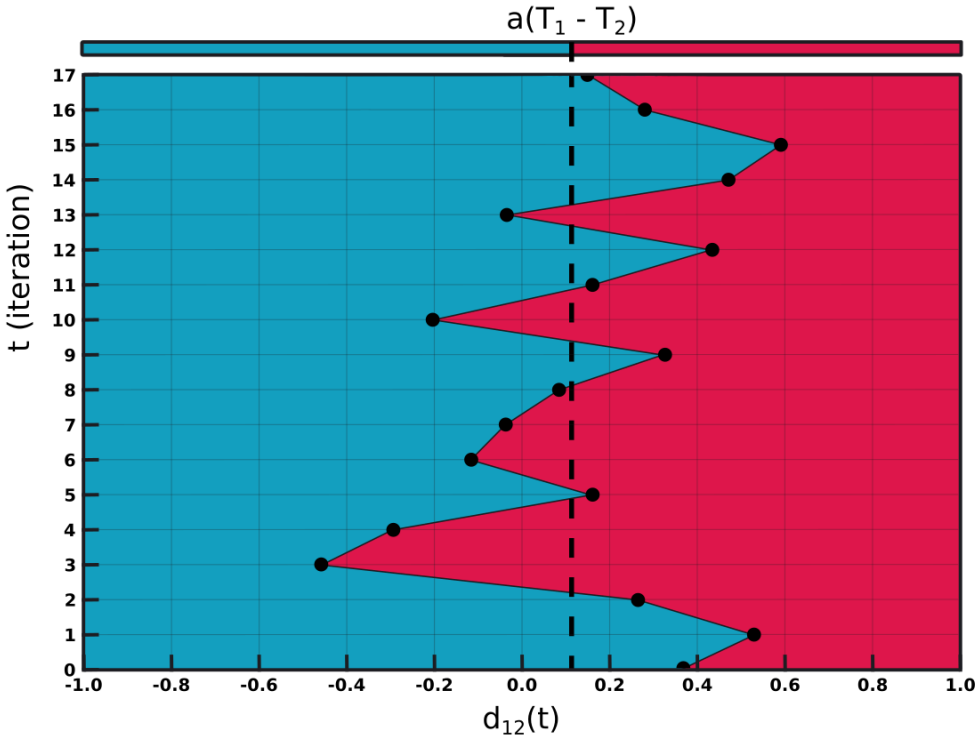

Figure 2. Example of the values given by the distance in terms of performance $d_{1,2}^{[a,\sigma_t,\sigma_c]}(t)$, defined by Equation (2), for $a(T_1 - T_2) = 0.11$ in 18 iterations t.

The placement of the $d_{1,2}^{[a,\sigma_t,\sigma_c]}(t)$ center is important because it determines the advantage or disadvantage of the agent with the label 1 over the agent with the label 2, the effect of which is mitigated or enhanced by the other parameters.

The quantities $\widetilde{P}^{[a,\sigma_t,\sigma_c]}(t)$ obtained from $d_{1,2}^{[a,\sigma_t,\sigma_c]}(t)$ have values defined between 0 and 1, and for the two agents involved in the match, they assume specular values with respect to the value 0.5.

2.3. The Tennis Agent-Based Model

Using Equation (1), it is possible to assign points to the agent-players at each iteration t, on the basis of the greater value of the performance between $\widetilde{P}_1^{[a,\sigma_t,\sigma_c]}(t)$ and $\widetilde{P}_2^{[a,\sigma_t,\sigma_c]}(t)$.

Having a way to assign the points in a simulated match provided us with the possibility to replicate the structure of rules of a real tennis match, and so obtain a virtual tennis match with a winning agent and a losing agent.

It was also possible to reproduce the structure of a tournament, and thus obtain the ranking of the agents playing in the tournament and the relative number of scored points, so as to obtain data comparable to the real ones.

The structures described above, even if they cannot be expressed analytically, are essential to reproducing the phenomena studied. They are complementary to the Equation (1) and modify the way these equations act to determine the evolution of the modeled system.

Figure 3 shows a scheme of the virtual match and the virtual tournament simulated by the agent-based model in which the agents compete. The simulated tournaments were constructed with a total of 128 agent-players, a number close to the average number of players that participate in an ATP tournament, and each simulated tournament had new

agents with a talent drawn from a predetermined talent distribution. The initial pairings between agents were random, and the subsequent ones were dictated by the winning agent-player. The simulated tournaments were composed of 7 rounds, with the winner of the final simulated game being the tournament winner.

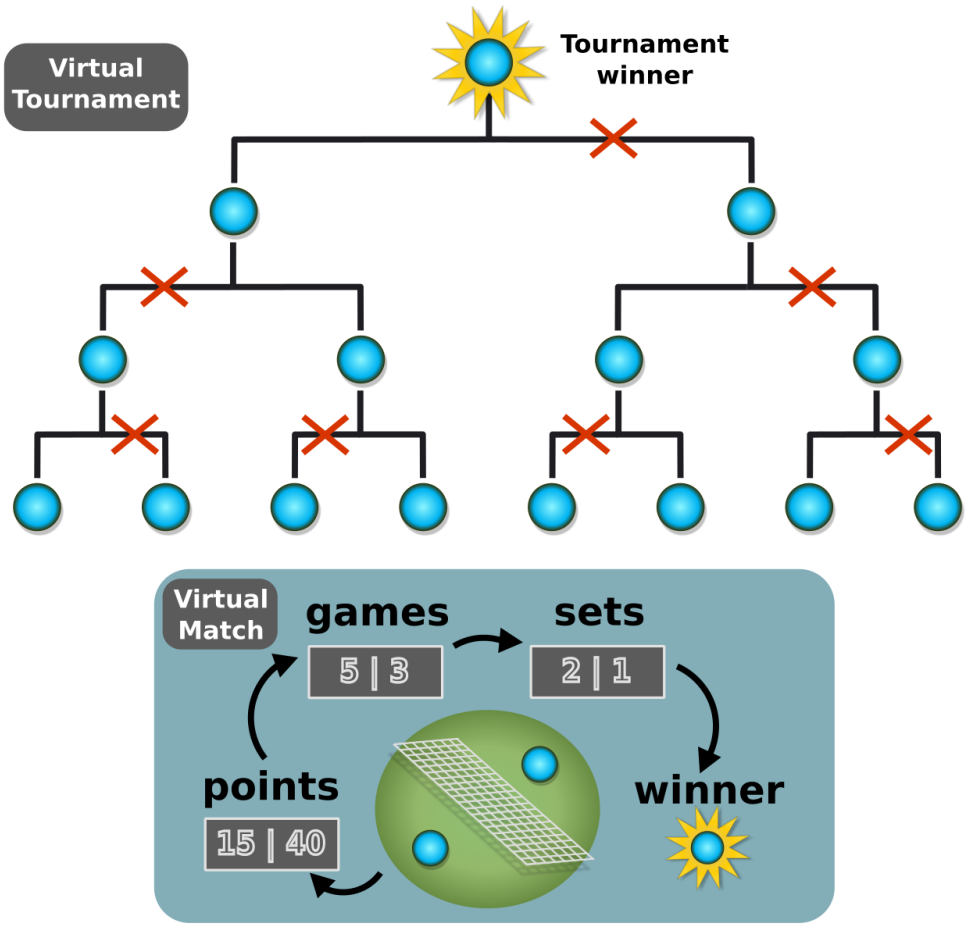

Figure 3. (**Bottom**) Diagram summarising the virtual match steps that determine the winning agent of the individual matches. (**Top**) Diagram showing how the virtual tournament determines a tournament winner based on the results of individual matches and subsequent matches in each tournament stage.

2.4. The Agent-Based Model Genetic Algorithm Calibration Setup

The calibration of our agent-based model was carried out using a genetic algorithm. This type of algorithm is often used in optimization problems [25], including with agent-based models [28]. In particular, the use of a genetic algorithm allowed us to explore the entire 3D parameter space of the agent-based model in search of parameter values that cause the simulated results to adhere to the real data.

The genetic algorithm was implemented using PyGAD, a library written in Python for the design of genetic algorithms. These algorithms attempt to simulate a natural selection process for the possible solutions of a model. Genetic algorithms use a function, the fitness function, to compare models with different parameter values, This function provides us with a fitness value that is higher for the models that are able to better adhere to the observed data with the set of examined parameters, called genes. Often, the models being compared are also simple functions, but in our case, the model being compared was an agent-based model that produces data after a certain number of simulated tournaments. Therefore, we needed to adapt the genetic algorithm to this type of model; specifically, the data obtained from the agent-based model could be organized in a score performance

distribution, as shown in Figure 1. The fitness function we required was one that allowed us to compare the real score performance distribution with the simulated score performance distribution obtained for a given combination of parameters; a good function for our needs was the inverse of a generalized Euclidean distance:

$$\begin{aligned} d(p,q) &= \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_n-q_n)^2} \\ f(p,q) &= 1/d(p,q) \end{aligned} \tag{3}$$

where $p_1, p_2 \ldots p_n$ and $q_1, q_2 \ldots q_n$ refer to the abscissae of the two normalized histograms representing the compared score performance distributions, $d(p,q)$ is the generalized Euclidean distance, and $f(p,q)$ is the fitness function.

Using this fitness function, the genetic algorithm optimizes the value of three genes, which are the three parameters of the model. For all three genes, the range of possible values is between 0.01 and 1.00. These values can be varied with steps of at least 0.01, so the algorithm is computed on a 3D lattice of 0.01 steps; this implies 1,000,000 possible combinations of the parameters.

In this scheme, for a particular set of parameter-genes chosen to be analyzed by the genetic algorithm, we needed to generate a score performance distribution to apply our fitness function. In order to optimize the calculation of the performance distribution, which involved a large number of tournament simulations, we set out a system to calculate the fitness with a variable amount of data. This involved a system of thresholds for the fitness values, with the first threshold being placed for a value of fitness equal to 0.10, and the successive steps of the threshold systems were spaced at 0.10 intervals, so the second threshold was 0.2, the third was 0.3, and so on. For the calculation of the first threshold, we used the data of only 100 simulated tournaments, and for each threshold step, we added 100 simulated tournaments. With this calculation scheme, we avoided wasting time in the calculation of models with parameters that provide low fitness values, and instead we calculated the models with parameters that provide high fitness values with satisfactory statistics.

In our setup, the algorithm starts its optimization process with a population, the number of studied triads of different parameters, of 30 individuals, and the algorithm iterates its optimization cycle for 100 generations. This approach allowed us to obtain a higher density of points in the parameter space for the region where the solutions with higher fitness values were located and a coarser density of points where there were solutions with low fitness values.

3. Results

3.1. The Agent-Based Model Simulation

The developed agent-based model was constructed using Equation (1), and so it depends on the same three parameters, the talent weight a, the standard deviation of the distribution of talent σ_t, and the standard deviation of the distribution of chance σ_c. Having chosen the values of the three parameters, the model runs a series of virtual tournaments, and we collect the data obtained from the virtual tournaments to construct, as in the real case, a score performance distribution.

Figure 1c shows the distribution obtained considering 10,000 simulated tournaments, each with 128 agents-players participating, with the fixed parameters a = 0.3, σ_t = 0.2, σ_c = 0.2. The distribution obtained from the model has the same characteristics as those obtained from the real data, and its shape depends on the three parameters of the model, so with the right combination of parameters, it is possible to obtain a distribution that fits the real distribution. An interesting feature observed in the simulations is that for a value of a equal to 1, representing pure talented-based tournaments, the area of intersection between the losing and winning parts of the score distribution disappears, so the presence of this intersection is due to the action of chance.

With the same set of parameters used for the distribution in Figure 1c, Figure 1d instead shows the variation of the FWHM at different stages of the tournaments, calculated in the same way as for real data. We can see that the model is able to reproduce the same kind of trend for the FWHM as shown for the real data. This trend, which has also been shown in other works [11–13], is due to the fact that the most talented players are selected to participate in the later stages of the tournaments.

3.2. Calibration of the Agent-Based Model on the Real Data

The calibration of the agent-based model with the help of a genetic algorithm is performed considering only the comparison between the real and simulated score performance distribution. We collected the data from 7 runs of the genetic algorithm, each starting with a population of 30 individuals and lasting for 100 generations, as described in previous sections.

Figure 4 shows the result of the calibration, in which the points of the parameter space with higher fitness values are distributed on a thin surface, in the layer with the lighter color, so the good parameters for the model are limited to the points on this surface.

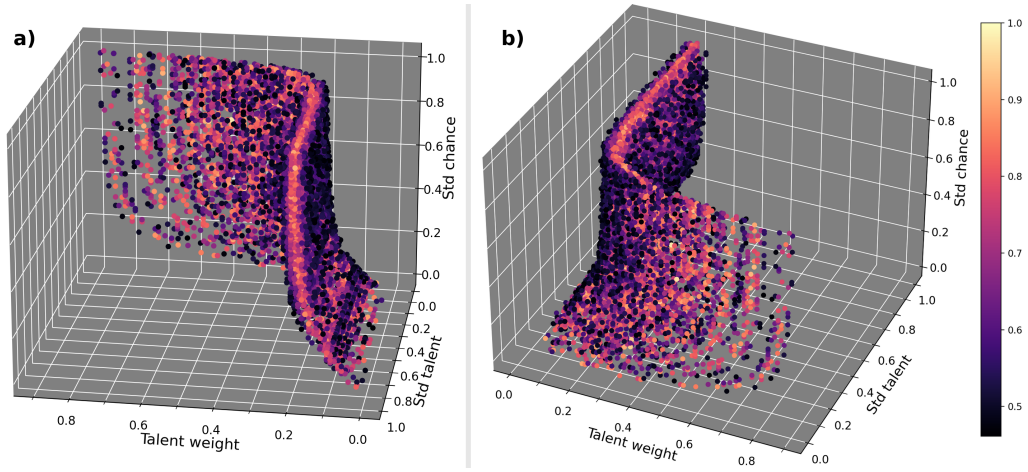

Figure 4. Results of the genetic algorithm calibration of the agent-based model. (**a**,**b**) Two perspectives of the same results, providing a better understanding of the 3D distribution of the points. In the graph, the points represent a triad of parameters, namely the talent weight a, the standard deviation of the talent distribution σ_t, and the standard deviation of the chance distribution σ_c, while the colors represent the normalized fitness value relative to the points. Only the points with a fitness greater than 0.46 are shown.

These results obtained using a genetic algorithm allowed us to locate an area of interest, a surface, in the parameter space. With this more circumscribed area, we can further investigate the ability of the agent-based model to replicate the real data; in fact it is now possible to confront the trend of the FWHM of the real data with that emulated by the agent-based model with parameter values belonging to this limited area.

3.3. Further Calibration of the Model on the Surface of Interest

Even if we restrict the region of interest, performing this additional confrontation for all points with high fitness values would be very costly in terms of computational efficiency. Therefore, we performed a simple fit of the region of interest and selected 1450 points appropriately distributed on the fitted surface, hereafter referred to as the surface of maximum fitness, thus limiting the computational complexity of the problem.

It is then possible to run the agent-based model for each of these 1450 points, which are parameter triads, simulating 7000 tournaments for each point, and thus obtaining for each point a simulated trend for the FWHM, similar to that shown in Figure 1d, which we

compare with the FWHM trend of the real data shown in Figure 1b, using the same $f(p,q)$ shown in Equation (3).

Thus, for each point, we obtain a sort of fitness value for the two FWHM trends, simulated and real. This quantity, which will be referred to as *"verisimilitude"* from here on, so as not to confuse it with the fitness used in the context of the genetic algorithm, is greater when the two trends being compared are similar.

Figure 5 shows the result obtained by using the method described above. The color map used in the Figure 5 allows us to visually identify the areas with the highest verisimilitude. Not all areas are equally highlighted by the color map, and two regions stand out among the points in the parameter space that seem to possess high values of verisimilitude. Table 1 shows the weighted mean values, in terms of verisimilitude, of the parameters in the two regions mentioned above; these values were calculated taking into account only the parameters with a normalized verisimilitude higher than 0.774. The values in green in the table correspond to an area with low values of a and σ_c, while the values in red in the table correspond to the area with high values of a and σ_c. The σ_t values are less variable in the two areas, but the parameters in the green area generally have higher values than those in the red area, and the red area in particular can have very low values.

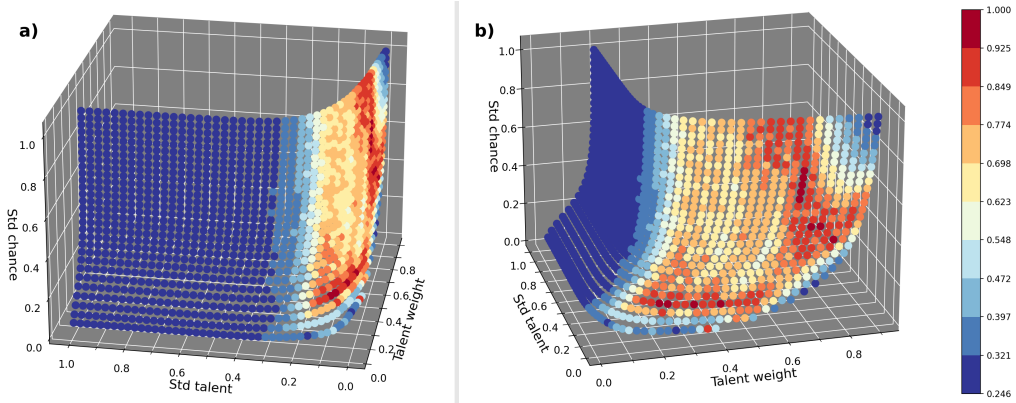

Figure 5. (**a**,**b**) Two different perspectives of the results. The color map highlights the areas that better match the real data with higher values of normalized verisimilitude. Only a limited number of the 1450 points have a high value of verisimilitude.

Table 1. Average values, weighted by verisimilitude, of the parameters for the two regions, with high verisimilitude values observed on the surface of maximum fitness in Figure 5. Only the points with a normalized verisimilitude greater than 0.774 are used to calculate the mean values of the parameters. The values in green are obtained for the regions with low values of σ_c, while the values in red are obtained for the regions with high values of σ_c.

	a	σ_t	σ_c
w. mean	0.34	0.07	0.16
w. mean	0.66	0.04	0.67

These weighted mean values represent in a compact way the properties of the parameters with a high verisimilitude in the two regions. The differences in the values obtained for these two areas lead us to carry out further analysis.

3.4. Calibration on the Final Phases of the Tournaments

The agent-based model constructed is deliberately simple, built using a limited number of parameters. However, for the initial phases of real tournaments, there may be several discrepancies between the different tournaments, concerning, for example, the fact that different tournaments in reality involve a different number of stages to reach the final. There may also be differences in the way that participants are selected. Thus, in the early

stages, it is possible to have very different distributions of talent among the participating players in different tournaments.

Our model does not predict these discrepancies; there may also be other factors, in addition to those given as examples, that are not taken into account and therefore not reproduced by this model. However, these discrepancies and differences from an ideal model tend to diminish in the more advanced stages of tournaments, because tournaments have a selection capacity that tends to mitigate any inhomogeneities in the talent distribution of the participants. It is therefore expected that in the final stages of tournaments, it will be easier to adapt the agent-based model to the real data.

Comparing only the tournament phases from the round of 16 onwards, and thus the last 4 tournament phases, in the same manner as in the previous section, considering 7000 simulated tournaments for each of the 1450 points on the surface of maximum fitness, we obtain what is shown in Figure 6.

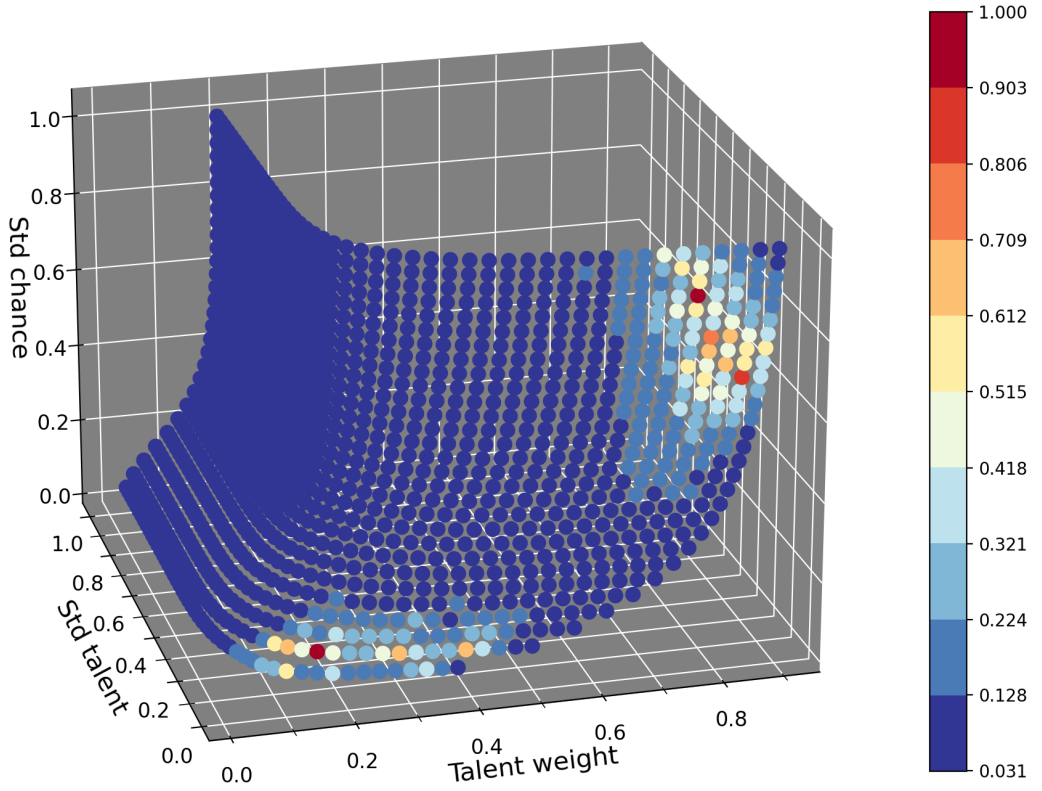

Figure 6. Result of the comparison between the real and simulated FWHM trends in the 1450 points of interest, considering only the last four tournament phases; the color map of normalized verisimilitude highlights the areas that better match the real data.

It can be observed that there are only a limited number of points in the parameter space that occupy the parts with the high verisimilitude limit, so only a limited number of points can be identified that are closer to the real trend in the later stages of the tournaments.

Table 2 shows the weighted average of the parameters and the parameters with the the highest verisimilitude for the two separate areas on the surface of maximum fitness observed in Figure 6. For the highest-verisimilitude parameters highlighted in green, we have $\sigma_c = 0.08$, $\sigma_t = 0.09$ and a talent weight $a = 0.17$; for this point, the model predicts that chance plays a role dictated by small fluctuations, given the limited amplitude of the distribution corresponding to the value of σ_c considered, but this is very relevant given the low value of the talent weight a, with players having a narrow talent distribution.

Table 2. Parameter values of the two points with the highest verisimilitude and the weighted average of the points with normalized verisimilitude greater than 0.321 for the two regions on the surface of maximum fitness, as highlighted in Figure 6 with high verisimilitude. The points in green in this table belong to the region with a low value of σ_c, while the red points belong to the region with a high value of σ_c.

	a	σ_t	σ_c
max ver.	0.17	0.09	0.08
w. mean	0.23	0.07	0.07
max ver.	0.78	0.03	0.90
w. mean	0.80	0.02	0.79

For the highest-verisimilitude parameters highlighted in red, we have $\sigma_c = 0.90$, $\sigma_t = 0.03$ and $a = 0.78$. So, lower values of σ_t characterize this point. The chance distribution is wider, meaning that random events that could change the fate of the matches are more frequent, but since the talent weight a is high, much higher than the point examined above, chance has less weight in determining the outcome of the matches.

3.5. Calibration by Parameter Constraint

The study of the real data with the agent-based model gives us two possible results, or rather two possible interpretations of the data; we find two areas of interest on the surface of maximum fitness. We can compare our results with previous work to understand which of the two interpretations of the data is an artifact of the model and which reflects the real situation.

For both model calibrations carried out in the previous sections, the region on the surface of maximum fitness with low values of σ_c has many parameters (for example in Table 2, we get $\sigma_t = 0.09$ for the maximum verisimilitude results), providing a talent distribution with an amplitude similar to that used in other works [11,12], a normal distribution with $\sigma_t = 0.1$, and a mean $\mu = 0.6$.

This talent distribution is derived from the population IQ distribution and thus comes directly from real data, so it makes sense to use this talent distribution in our model to solve the double interpretation problem. Then, considering the green values in Tables 1 and 2, we can also fix σ_c by considering the average of the σ_c values, obtaining $\sigma_c = 0.12$ as the best representative value.

Thus, by setting $\sigma_t = 0.1$ with $\mu = 0.6$ and $\sigma_c = 0.12$ in our agent-based model, it is possible to again calibrate the model, minimizing the distance between the data reproduced by the model and the real one, using the same method as in the previous sections, with only a as a free parameter.

Figure 7 shows the result of the calibration. The normalized verisimilitude value shown is obtained by simultaneously taking into account the score performance distribution and the trend of the FWHM, comparing them with the real data through $f(p, q)$ shown in Equation (3), and then taking into account the normalized average value.

There are three talent weight values a with a normalized verisimilitude greater than 0.85, as shown in Figure 7. Thus, fixing the values of the other two parameters, with plausible values suggested by the free parameter model of the previous sections and the literature, the value of the talent weight a for tennis competitions on the ATP tour is between 0.18 and 0.20. These values are in good agreement with other works on the subject [13].

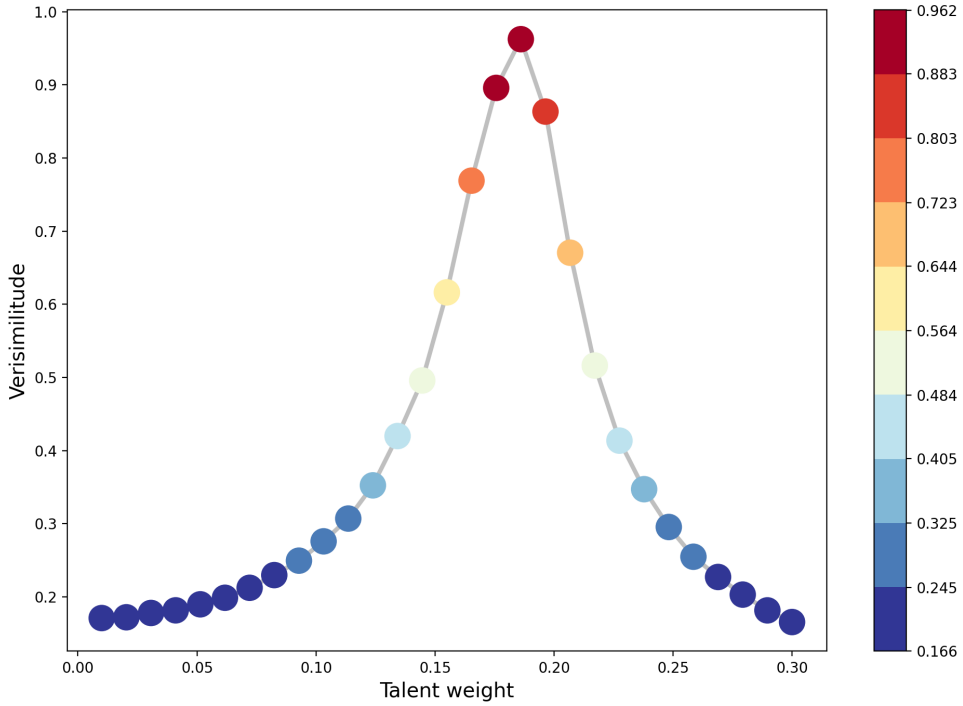

Figure 7. Result of the calibration of the agent-based model with the single talent weight parameter a. The x-axis presents the values of a, and the y-axis provides the corresponding values of verisimilitude.

4. Discussion

In this paper, we have studied the impact of random events, i.e., the action of chance, as opposed to the role of talent in determining the outcome of a direct competition; in particular, we have studied tennis competitions, and the data obtained from ATP tournaments.

The aim of this work was to obtain as much information as possible about this type of competition, with the fewest assumptions about the parameters in the model, in order to obtain the widest range of information. To achieve this, we have developed an agent-based model capable of reproducing the main features of the data, in particular the score performance distribution and the trends of the FWHM of these distributions over the different phases of the tournaments. The agent-based model developed depends on three parameters, namely the talent weight a, the standard deviation of the talent distribution σ_t, and the standard deviation of the chance distribution σ_c, so we made the assumption that talent and chance can be described by a normal distribution centered at 0.5. Then, we used a genetic algorithm to analyze the parameter space, and thus considered all the possible combinations of parameters of the model able to make the simulated score performance distribution similar to the real one. The use of a genetic algorithm allowed us to make very few assumptions about the values of the parameters and thus allowed us to find a surface, the surface of maximum fitness, in which the points, corresponding to the parameter triads, are able to reproduce score performance distributions very similar to the real ones.

Considering then only a limited set of 1450 points, uniformly distributed on this surface of maximum fitness, we carried out a further calibration of the agent-based model comparing another feature of the data, the trend of the FWHM for the different phases of the tournaments. This allowed us to further restrict the set of parameters that are able to reproduce the trend of the real data. We performed this further calibration by first considering all phases of the tournament, and then considering only the last four phases of the tournament, in an attempt to mitigate the effects of factors not considered by the agent-based model.

In both cases, we found a limited set of good points able to reproduce the real data. These points are distributed in two regions of the surface of maximum fitness—the regions with high values of σ_c, characterized by higher values of talent weight a and lower values

of σ_t, and the regions with low values of σ_c, characterized by lower values of a and higher values of σ_t. The most interesting fact about these two regions is that they underline a sort of redundancy in the equations that govern the agent-based model. In fact, it seems that for the model, a and σ_c are two parameters able to counterbalance their effects, as high values of talent weight a correspond to low value of chance weight, so we found a region with high values of chance weight but a small amplitude of the chance distribution and a region with low values of chance weight but a large amplitude of the chance distribution.

The agent-based model and the genetic algorithm therefore give us two areas of interest on the surface of maximum fitness, but only one of these two regions contains parameters that correspond to the real situation, while the other one is given by this sort of redundancy in the model. The two areas have talent distributions of different amplitude, providing us with the ability to choose one of the two areas by selecting a specific talent distribution. So, we choose the area on the surface of maximum fitness with parameters having a low value of σ_c and slightly higher values of σ_t than the other area of solutions. This choice was made by comparing our results with previous works [11–13]: the parameters in the low σ_c area have σ_t values similar to the talent distribution used in the cited works, extrapolated from the IQ distribution, and thus obtained from the real data.

Using the IQ talent distribution and a $\sigma_c = 0.12$ obtained considering the average of our solutions in the low σ_c area, we carried out a further calibration of the agent-based model and obtained a value talent weight between $a = [0.18, 0.20]$. This value is in good agreement with previous works on the subject [11–13]. This result gives us confidence that the solutions in the low σ_c region of the surface of maximum fitness are those to be considered. The solution given in Table 2 with maximum verisimilitude in this region, obtained by our agent-based model trained with a genetic algorithm, has $a = 0.17$, $\sigma_t = 0.09$, and $\sigma_c = 0.08$. These values seem reasonable; in fact, with σ_c in this range, we expect that the random events that occur in the match will mostly be small, such as wind, which can change the trajectory of the ball, uneven ground, etc., and the events that can change the outcome of the match will be rare, such as an injury. With σ_t in the predicted range, we expect the majority of players to have similar talent levels, with only a few individuals having talent levels that are higher or lower than those of others. Last but not least, for the talent weight a in the predicted range, we expect the action of chance to play a large role in determining the outcome of a match. However, this does not mean that talent does not matter, as previous works have shown [5,9,12,13], but that in a match between two individuals of similar talent, chance plays a major role, confirming the so-called *talent paradox* [13], and we expect the outcome to be in favor of one rather than the other by sheer luck. Talent becomes more important as the difference in talent between the two competitors increases, but even when mitigated, chance cannot be ignored. Our model therefore predicts a scenario in which there are many random events, most of which are small, but which become important when the talent levels of the players are comparable.

5. Conclusions

The goal of this work was to obtain information about tennis competitions without major a priori constraints on the quantities that determine the outcomes of these events.

We have shown that with an agent-based model constructed to emulate the data of tennis competitions, a genetic algorithm that helps us to explore the parameter space, and related analysis, it is possible to infer the range of values that the parameters a, σ_t and σ_c can take, providing very useful information about this type of competition. We have confirmed the importance of the action of chance for these competitions, finding that with both fixed-parameter and free-parameter calibration approaches, chance has a weight around 80% in determining the outcome of a tennis competition, with narrow distributions of talent and chance, having standard deviation values around 0.01.

This type of approach is generalizable to other types of competitions provided that we are able to construct an appropriate agent-based model and that we have a good amount of data about the competition under study to calibrate the model.

We were also able to better understand the sensitivity of this type of model to the values of the parameters, and we have highlighted the possibility of a kind of redundancy, and so highlight the importance of choosing the right values for the parameters. The choice of the chance distribution with the right characteristics seems to be very important in this type of model; in fact, the talent weight and the chance distribution seem to have a complementary role.

To resolve this kind of redundancy, it is necessary to impose some kind of constraint on the model, based on real data and experimental evidence. In future work, it would be interesting to understand whether this redundancy can be eliminated with some modification to the core model equations or to the rule structure of the agent-based model. Another measure that could be considered is using distributions with unfixed centers of symmetry for the talent and chance distributions, although this would make the model depend on five parameters instead of three, making convergence to the real data more difficult.

Even without these modifications, we have established a useful method that is applicable to a wide range of competitions involving individuals or entities competing against each other, extrapolating quantitative information regarding the role of chance and randomness in the final outcomes.

Author Contributions: Conceptualization, S.P. and A.R.; software and numerical simulations, S.P.; validation, S.P. and A.R.; writing—review and editing, S.P. and A.R.; supervision, A.R.; funding acquisition, A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the the project PRIN 2017WZFTZP "Stochastic Forecasting in Complex Systems".

Data Availability Statement: The data used are available online, see refs. [22–24].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nisperuza, J.; Rubio, J.P.; Avella, R. Density probabilities of a Bose-Fermi mixture in 1D double well potential. *J. Phys. Commun.* **2022**, *6*, 025004. [CrossRef]
2. Meerson, B.; Sasorov, P.V. Domain stability, competition, growth, and selection in globally constrained bistable systems. *Phys. Rev. E* **1996**, *53*, 3491–3494. [CrossRef] [PubMed]
3. Miao, R.; Chun, H.; Feng, X.; Gomes, A.C.; Choi, J.; Pereira, J.P. Competition between hematopoietic stem and progenitor cells controls hematopoietic stem cell compartment size. *Nat. Commun.* **2022**, *13*, 4611. [CrossRef] [PubMed]
4. Metcalfe, J.; Ramlogan, R.; Uyarra, E. *Economic Development and the Competitive Process*; Centre on Regulation and Competition (CRC) Working papers 30612; University of Manchester, Institute for Development Policy and Management (IDPM): Manchester, UK, 2002.
5. Rapisarda, A.; Pluchino, A.; Biondo, A.E. Talent Versus Luck: The Role of Randomness in Success and Failure. In *Advances in Complex Systems (ACS)*; World Scientific: Singapore, 2018. [CrossRef]
6. Barabási, A.L. Untangling performance from success. In *EPJ Data Science*; Springer: Berlin, Germany, 2016. [CrossRef]
7. Sinatra, R.; Wang, D.; Deville, P.; Song, C.; Barabási, A.L. Quantifying the evolution of individual scientific impact. *Science* **2016**, *354*, aaf5239. [CrossRef] [PubMed]
8. Fraiberger, S.P.; Sinatra, R.; Resch, M.; Riedl, C.; Barabási, A.L. Quantifying reputation and success in art. *Science* **2018**, *362*, 825–829. [CrossRef] [PubMed]
9. Pluchino, A.; Burgio, G.; Rapisarda, A.; Biondo, A.E.; Pulvirenti, A.; Ferro, A.; Giorgino, T. Exploring the role of interdisciplinarity in physics: Success, talent and luck. *PLoS ONE* **2018**, *14*, e0218793. [CrossRef] [PubMed]
10. Pluchino, A.; Rapisarda, A.; Sinatra, R.; Zappalà, C.; Sousa, S.; Cunha, T. Early Career Wins and Tournament Prestige Characterize Tennis Players' Trajectories. In *EPJ Data Science*; Springer: Berlin, Germany, 2024. [CrossRef]
11. Rapisarda, A.; Sobkowicz, P.; Frank, R.H.; Biondo, A.E.; Pluchino, A. *Inequalities, Chance and Success in Sport Competitions: Simulations vs Empirical Data*; Elsevier B.V.: Amsterdam, The Netherlands, 2020.
12. Zappalà, C.; Pluchino, A.; Rapisarda, A.; Biondo, A.E.; Sobkowicz, P. On the role of chance in fencing tournaments: An agent-based approach. *PLoS ONE* **2022**, *17*, e0267541. [CrossRef] [PubMed]
13. Zappalà, C.; Rapisarda, A.; Biondo, A.E.; Pluchino, A. The Paradox of Talent: how Chance affects Success in Tennis Tournaments. *Chaos Solitons Fractals* **2023**, *176*, 114088. [CrossRef]
14. Fink, T.M.A.; Coe, J.B.; Ahnert, S.E. Single elimination competition. *Europhys. Lett.* **2008**, *83*, 60010. [CrossRef]
15. Ben-Naim, E.; Hengartner, N.; Redner, S.; Vazquez, F. Randomness in Competitions. *J. Stat. Phys.* **2013**, *151*, 458–474. [CrossRef]

16. Salgado, M.; Gilbert, N., Agent Based Modelling. In *Handbook of Quantitative Methods for Educational Research*; Teo, T., Ed.; SensePublishers: Rotterdam, The Netherlands, 2013; pp. 247–265. [CrossRef]
17. Bak, P.; Paczuski, M. Why Nature is complex. *Phys. World* **1993**, *6*, 39. [CrossRef]
18. Cenani, S. Emergence and complexity in agent-based modeling: Review of state-of-the-art research. *J. Comput. Des.* **2021**, *2*, 1–24. [CrossRef]
19. Hawick, K.A. *An Agent Model Formulation of the Ising Model*; Technical Report; Information and Mathematical Sciences, Massey University: Auckland, New Zealand, 2003.
20. Sznajd-Weron, K.; Jędrzejewski, A.; Kamińska, B. Toward Understanding of the Social Hysteresis: Insights from Agent-Based Modeling. *Perspect. Psychol. Sci.* **2023**, *19*, 511–521. [CrossRef] [PubMed]
21. Levayer, R. Cell competition: Bridging the scales through cell-based modeling. *Curr. Biol.* **2021**, *31*, R856–R858. [CrossRef] [PubMed]
22. ATP Tour Site, Info and Statistics about Tennis. Available online: https://www.atptour.com/en/ (accessed on 30 June 2023).
23. ATP Tour Site, for Tennis Data. Available online: https://datahub.io/sports-data/atp-world-tour-tennis-data (accessed on 30 June 2023).
24. JeffSackmann Github Repository of Tennis Data. Available online: https://github.com/JeffSackmann/tennis_atp (accessed on 30 June 2023).
25. Lingaraj, H. A Study on Genetic Algorithm and its Applications. *Int. J. Comput. Sci. Eng.* **2016**, *4*, 139–143.
26. Tennis Rules and Info Site. Available online: https://olympics.com/en/news/tennis-rules-regulations-how-to-play-basics (accessed on 30 June 2023).
27. Tennis Rules Site. Available online: http://protennistips.net/tennis-rules/ (accessed on 30 June 2023).
28. Joyce, K.E.; Hayaska, S.; Laurienti, P.J. A genetic algorithm for controlling an agent-based model of the functional human brain. *Biomed. Sci. Instrum.* **2012**, *48*, 210.