University of BRISTOL

Department of Life Sciences

BIOLM0028 Thesis

09/2022

A MODEL COMPARISON APPROACH FOR THE COLLECTIVE DECISION-MAKING OF FISH

Student ID: 2150354

MSc Bioinformatics Thesis

Word Count: 5967

By submitting this assignment cover sheet, I confirm that I understand and agree with the following statements:

- 'I have not committed plagiarism, cheated or otherwise committed academic misconduct as defined in the University's Assessment Regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)
- 'I have not submitted this piece, in part or in its entirety, for assessment in another unit assignment (including at other institutions) as outlined in section 4 of the University's Assessment regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)
- I understand that this piece will be scrutinised by anti-plagiarism software and that I may incur penalties if I am found to have committed plagiarism, as outlined in sections 3 of the University's Examination Regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)

I give permission for my work to be used anonymously in examples given to students in the future

ABSTRACT

Individuals in animal aggregates use social cues to make decisions on movement direction. Recent models have attempted to determine the types of social information used in these mechanisms. However, the large number of theoretical models being proposed has led to some uncertainty on the importance and weight of these determinants in decision-making. This study will help assess previously identified determinants for the decision-making of Trinidadian river guppies (Poecilia reticulata) using a model comparison approach, while providing novel models for animal decision-making. Shoals of 8 fish were allowed to freely roam a y-maze in a non-manipulative approach. Using this approach, events of binary decision-making (swim in the arm on the right, or on the left) naturally occurred. Using logistic regression models fit with social information data from either arm, we note that group size is an important social cue in continuous decisions (when a fish must decide whether to follow a cohesive shoal). When it comes to binary decisions (when a fish must decide between two separate shoals) we instead confirm the importance of distance as a decision-making determinant, where fish are more likely to choose an option if the closer shoal, or their nearest neighbour, has chosen that option. We also discuss the possibility of fish using the last observed choice to make their decision in cases of minimal availability of other social cues. Finally, this study introduces two novel determinants for binary decision-making: the speed of a shoal (fish were more likely to follow the quicker shoal, given that the shoal is nearby), and the group size of each shoal accounting for average orientation (fish were more likely to follow the larger number of fish in each shoal which match the swimming direction of the focal fish).

Keywords: Collective behaviour, fish shoals, decision-making, orientation, metric range, topological range, quorum responses, social animals, statistics, logistic regression

Dedications & Acknowledgements

I would like to thank my supervisor, Dr. Christos C. Ioannou, for helping me throughout the process of writing this thesis, as well as my academic advisors, Dr Celine Petitjean and Dr Jordi Paps Montserrat, for organizing a great Masters course.

I dedicate this study to my family, who has always supported my choices and work. I also acknowledge the support many friends have given me, particularly those who have helped me develop my academic skills and who have inspired me to pursue this academic career.

CONTENTS

1 INTRODUCTION

Collective animal behaviour is a phenomenon in biological systems based on the coordination of social animals to self-organize [1, 2] in order to ensure safety in the context of predator evasion [3] and better success in foraging [4]. The methods adopted to study these dynamics include: theoretical models used to establish the dynamics of social living (leadership mechanisms, cohesion, etc.) [5,6]; manipulation of sensory information to test the response of animals to their environment [7, 8]; the social movements of fish at the individual level, highlighting the costs and benefits of social living [9, 10]; and statistical models fit to lab movement data to estimate the nature of collective and individual decisions with the end goal of understanding the self-organization of animal aggregates [11, 12]. This last method has attempted to provide evidence for the interaction rules responsible for the decision-making of animals, but the generic nature of these models has made the selection of the best theoretical assumptions quite complex [13].

Fish in particular have been extensively studied due to their innate and highly organized shoaling capabilities, which lead to more cohesive shoals [14]. Some of the most popular determinants of fish decision-making in the literature are based on the spatial positioning of each fish compared to the rest of the animal aggregate [15, 16]. The main models based on spatial positioning are known as metric range models. In theory, animals can determine the distance between each other, and thus only the neighbours within a fixed proximity are believed to affect the decisions of a focal fish. These decisions are not believed to be instantaneous, but rather hierarchical, meaning fish will assess the presence or absence of one or more neighbours in zones of attraction, repulsion or neutrality [12, 17, 18]. These three zones are believed to create an equilibrium where fish will not fall behind the shoal by following their neighbours (attraction) but will also not get too close to their neighbours, therefore avoiding collision (repulsion). Evidence for accurate distance estimations in fish makes these metric interactions seem natural [19]. Another explanation is that of quorum responses, which are defined as a steep increase in the probability of a focal fish performing a given behaviour once a minimum threshold number of neighbouring fish already performing that behaviour is exceeded [20, 21]. While metric ranges may seem natural, they cannot reproduce the density changes typical of animal aggregations. Quorum responses, on the other hand, seem indispensable in maintaining cohesion during these density changes due to external stimuli, such as predation events [22] where animals simply using a metric range may lose track of their neighbours due to large-scale changes in distances. Some studies have also suggested that focal fish may follow the last observed choice from one of their neighbours [23].

The polarization of neighbouring fish, defined as the average deviation of each fish's orienta-

tion from the average direction of a shoal [24], is believed to play an important role in the self-organization of fish groups through alignment mechanisms. Neighbours therefore undergo parallel orientation in the previously mentioned neutral zones in order to maintain cohesion [24–26]. In a decision-making context, the positional and orientational mechanisms of a shoal simply act upon speed (a fish oriented towards another fish will cause the latter to slow down) [27], which in theory acts upon the turning tendencies of fish in order to control the decision-making of the following fish [28]. Orientation has therefore often been associated in decision-making as an intermediate to speed and with the end goal of parallel alignment, but has rarely been used as a stand-alone component of decision-making in animals. However, this evidence hints at a possible use for orientation in decision-making, as fish may be more likely to follow their neighbours if they follow the same direction, therefore possibly not only affecting alignment, but also being involved in creating zones of repulsion or attraction as identified in fish interaction modes [29].

The speed of animals in a social context has not been discussed extensively as a social factor for decision-making. Instead, studies have focused on the organizational importance of speed to ensure cohesion and effective evasion from predators [30]. However, higher average speeds in fish shoals usually suggest healthier fish [31] and, in many cases, suggest nearby food sources [4]. Assuming fish will most likely follow a healthier group with more successful foraging, higher speeds may be a determinant in decision-making, where focal fish will follow the quicker group on average. This may therefore introduce a novel interaction rule which has rarely been discussed in fish decision-making.

This study will use statistical comparisons of logistic regression models in order to determine which of these discussed social cues are most influential in fish decision-making. The use of a y-maze, along with free-roaming groups of eight guppies, will lead to decision-making events with binary options: left or right. This approach is non-manipulative, meaning the fish are free to roam through the y-maze without any external stimuli, and are required to make decisions where the three arms meet (decision triangle). Prior studies have used y-mazes as a means of minimising bias [3, 10], maintaining that decision will be based on social information alone. Because of this, fish are expected to have no preference for either arm in the absence of social information. Furthermore, when social information is present, fish are also expected to have no preference for either arm if the information in both arms is identical. Spatial positioning, group sizes, orientation and speed will be the main focus due to their established links in decision-making [10, 13, 23–29]. These four determinants are therefore all expected to perform better than randomness in predicting turning decisions. The data from these interaction rules will be used to train a machine learning (ML) classifier to predict the turning decisions of fish, where a good score will suggest a high likelihood that these interaction rules are what determines binary decision-making events in a social context.

2 MATERIALS & METHODS

This study focuses on the effect of social cues on the movement decisions of individual fish while exploring a novel environment. The environment was designed to determine individual decision-making as clearly as possible by identifying the centre of a radially symmetric y-maze, which was established as a “decision triangle”. The occurrence of movements from one area to another, while passing through the “decision triangle”, therefore marked a “decision-making event”. The fish undergoing a decision-making event, and therefore those assessing the social information gathered from other neighbouring fish, are referred to as focal fish. Tracking data was used to determine the positions of each individual fish at any given time-point. In all cases, there were 8 individual guppies involved in these movements. Groups used in each trial were all of the same sex and from the same site (thus exposed to the same level of predation) to account for group-level differences, as these can affect decision-making [32]. The tracking of fish positions was conducted as part of a previous study on the effects of predation on the social dynamics of collective exploration [3]; although this study only analysed the data from trials of 2 and 4 fish.

2.1 Filming & Tracking

Each arm of the y-maze was 36.5cm long and 10cm wide. The walls were 18cm high, while the water was 7.5cm deep. The arena was surrounded by plastic sheets to minimize lighting disturbance as it was filmed from above at 25 frames per second and a resolution of 1920 x 1080 using a Canon 550D DSLR camera. A removable door at the end of one arm was used to create a habituation area for the test subjects where they were stationed for 2 minutes before being released into the maze. The door was lifted at the start of each trial and the fish were recorded for 5 minutes. Each fish was only used in a single trial. The first 30 seconds of the video were not included in the tracking data to ensure all fish had an equal understanding of the y-maze. The source code for the tracking software is available at: https://github.com/ctorney/fishOfInterest. At each time-step (i.e. frame in the video), fish were allocated to an arm based on the coordinates (in pixels) of their position (0, 1, 2 or -1 for the decision triangle; see figure 1). In some cases, individuals disappeared from the tracking data due to occlusions between two or more fish, leading to instrumental errors. Once reappeared, a new unique ID is given to the individual. While these issues do not affect the positioning of fish, individual differences between the fish were not taken into account, due to problems with the identification of unique IDs. The collection of the fish and the trial procedure were carried out in March and April 2014 and were approved by the University of Bristol Ethical Review Group (UIN/13/028).

2.2 Predictor & Response Data Collection

The collection of social information data was carried out using the raw data from the tracking software. This approach involved the use of the Pandas 1.4.2 and Numpy 1.23.0 packages in Python 3.8. A fish crossing the decision triangle boundary (arm -1) represented the start of a decision event. Once this occurred, the arm the fish was leaving was ignored while the other two arms were set as right and left directional options. Social information data was collected from the right and the left arm. The response variable was the end of the decision, i.e. which arm the fish swam into. This was recorded as a binary response (either 0 or 1 for the right or left arm, respectively) which enabled the use of logistic regression models.

2.3 Tracking Errors

The fish entered and left the decision triangle 5857 times. However, only 3708 error-free decisions were used in this analysis. Following are the reasons for the exclusion of the remaining 2149 decisions. 1004 decisions: initial frame had less than eight fish due to tracking software issues. 312 decisions: missing neighbouring fish within the ending frame. 524 decisions: focal fish disappearing within a decision event. 231 decisions: previously missing fish reappeared within the decision triangle, and thus the initial position was unknown. 63 decisions: focal fish turned back to its initial arm (neither right nor left). 15 decisions: fish in the decision zone when the first 30 seconds of the trial had finished (initial position not recorded).

2.4 Statistical Analysis & Machine Learning

The focal fish decision on which arm to swim into (right or left, i.e. 0 or 1) is dependent on the right arm data (x_r) and the left arm data (x_l) as additive main effects (no interaction terms were included). The models were built in R 4.2.1 using general linear mixed models (GLMM) using the glmmTMB 1.1.4 package in R as binomial logistic regressions. We can produce a basic analytic expression for y* (log odds of P(L)) and, subsequently, an expression for both P(L) and P(R):

$$y = \beta_i + \beta_r x_r + \beta_l x_l$$
$$P(L) = \frac{e^y}{e^y + 1}$$
$$P(R) = 1 - \frac{e^y}{e^y + 1}$$

The sex of the fish was added as a main effect. The fish were collected from two sites for each of three Trinidadian rivers: lopinot, aripo and turure. In total, 42 five minute trials with 8 fish within the y-maze were conducted. Each site had multiple trials. Therefore, the trial number

was considered within the umbrella effects of the site as a nested random effect. Once all the models were built to predict the focal fish decision, the likelihood of the models given the data were compared using the Akaike Information Criterion corrected for small sample sizes (AICc) using the ICtab function in the bbmle 1.0.25 package. Further conclusions were drawn using statistical tests from the performance 0.9.2 (collinearity analysis) and the generalhoslem 1.3.4 (Hosmer-Lemeshow test) packages in R.

Using the scikit-learn 1.1.2 package on python, supervised learning classifiers were used to create ML models. These were then fit with the predictor data. In this study, the K-Nearest Neighbours, Decision Tree Algorithm and Naïve Bayes classifiers were trialled due to their common strengths in dealing with continuous explanatory data. Principal component analysis (PCA) within scikit-learn was conducted as a means of linear dimensionality reduction to project the data to a lower dimensional space and avoid overfitting. GridSearchCV within sckikit-learn was used to predict the best parameters for the algorithms.

3 RESULTS

3.1 Explanatory data collection

To account for and encompass most theoretical models that deal with counting data such as quorum decision models [20–22], the predictors NN (Number of Neighbours) for each option (i.e. turn right or turn left) were used as main effects to build two logistic regression models (for all decisions and for exclusively binary decisions). These predictors are a count of total fish in each option (i.e. right arm or left arm). The ”last observed choice” (o or 1, i.e. right or left), was instead used as the only main effect to build another model. This decision follows the assumptions of topological ranges, which assume that a focal fish will only follow N number of neighbours [13] (in this case, one neighbour).

To account for models based on distance, such as metric ranges [17,33], NND (Nearest Neighbour Distance) and AND (Average Neighbour Distance) models were built. To find these distances, the arm position for each neighbour at the end of the focal fish decision (i.e. when the focal fish exits arm -1) was extracted, as the final positions of all fish are of interest rather than the starting positions. The distances for the other fish in the two directional options were calculated using the difference between the focal fish coordinates and the neighbour coordinates at the start of the focal fish decision (i.e. when the focal fish enters arm -1), as we assume this is when the focal fish will commit to following a neighbour or group. NND was the shortest distance calculated in each option, while AND was a mean of all the distances for the shoals in each option.

Novel speed-based decision-making models stem from the evidence for speed as an intermediate to decision-making determinants [18, 34]. The speeds of each fish were measured using their coordinates at the start of a decision and one frame after the start. The distance between these two points was then divided by the time taken to swim this distance (i.e. one frame, which is 0.04 seconds). The first model was built using the nearest neighbour speed (NNS) in each option, which was identified by finding the nearest neighbour (using previously calculated NND values) and then selecting their speed. The second model was instead built using the mean speed of the shoals in each option.

Orientation has mostly been described as an important factor for the self-organization of fish through parallel alignment with near neighbours [24,30]. Here, orientation will instead be tested as a social determinant in decision-making. The first orientation model was built using the the mean angle that all fish create within their respective arm (ANO, Average Neighbour Orientation); in other words, how polarized each shoal is in each option relative to the swimming direction of the focal fish. This angle was calculated by using the slope and intercept of the

arm (each arm creates a line within the space) as well as the slope and intercept of the trajectories of each fish. These values range from 0° (opposite swimming direction to focal fish) to 180° (same swimming direction as focal fish). The second model was built using a group of predictors defined as the number of other fish swimming in the opposite direction to the focal fish (angle $< 90°$), and the number of other fish swimming in the same direction as the focal fish (angle $> 90°$) for each arm (NNO, Number of Neighbours Orientation) thus using counts rather than mean values.

3.2 Model Results

Of the 3708 Decision-Making Events (DMEs), 164 were leader events (i.e. no fish ahead of focal fish, lDMEs, 4.4%), 2252 had fish in either of the two options (i.e. oDMEs, 60.7%), and 1292 had fish in both options (i.e. binary decision-making events, bDMEs, 34.9%). lDMEs are instances where the focal fish does not have access to any social information due to being the first fish in the pack. In these events, the focal fish turned right in 49.9% of cases and turned left 50.1% of cases, showing a lack of directional bias, which may have been caused by behavioural lateralisation [35]. The main logistic regression models were built using bDMEs.

Two separate models could be fit to the counting (NN) data, the first using only bDMEs (figure 2A) and the second using all DMEs (figure 2B). The count predictors show good statistical significance in the first model ($p_{right} < .001$, $p_{left} = 0.008$) and very strong statistical significance in the second (p_{right} & $p_{left} < .001$). In 70.24% of all DMEs with unequal number of fish in each arm, the focal fish chose the most crowded arm. However, this value dropped to 60.3% using only bDMEs. These results, along with the coefficients in the model with all DMEs influencing the probabilities more, suggest avoidance for empty arms and an increased likelihood of following a shoal with an increase in shoal size. However, when at least one fish is present in the arm, choosing the more crowded arm is not as important. This is clear from the weaker predictive power of the model using only bDMEs.

The discrete nature of the ND variable created a simplified NND model which could perform better due to the straightforward nature of the data type. Being that the predictor is a binary variable, this model yields two conditional probabilities: the probability of turning left given that ND turned right ($P(L|R) = 0.387$) or left ($P(L|L) = 0.613$) (figure 2D). The odds ratio (OR) of the ND value shows an increase in likelihood that a fish will turn left if ND = 1 ($OR > 1$), and the variable is statistically significant ($p < .001$). Furthermore, the binary nature of the predictor allows us to fit this data to all DMEs, which provided us with two new conditional probabilities: $P(L|R) = 0.403$ & $P(L|L) = 0.597$. When comparing these predicted probabilities to those in the model using only bDMEs, there is a negligible difference.

In the NND model, the probability of turning left was at a minimum (P(L) = 0.061) when NND in the left arm was at a maximum (5cm), and at a maximum (P(L) = 0.973) when NND in the right arm was at a minimum (0cm) (figure 2C). When the right arm data and the left arm data are equal, the probability of turning left is close to 50% (P(L) = 0.53). Both predictors (right and left NND) show strong statistical significance (pright & pleft $<$ .001). In the average distance model, the statistical significance of the predictors is again very high (pright & pleft $<$ .001). The predictive capabilities are fairly strong (figure 2E) as seen from the large differences in OR values for right (OR = 1.921) and left (OR = 0.476) arm, which indicate that higher values for the right arm data have a strong positive effect on P(L), and vice-versa. However, the predictions are on average closer to 0.5, which suggest predictions with lower accuracy. This drop in the quality of predictions can also be seen from the coefficients for the right and left arm values of NND (table 1), which have a greater effect on the log odds compared to the coefficients for the right and left arm values of AND. There is no evidence to suggest areas of repulsion at shorter distances in either model.

Both ANS and NNS show little to no significance (p $>$ 0.05) as predictors in their respective models. Furthermore, their coefficients within the model hardly affect predictions. The logistic regression curves show a slight positive effect on P(L) when the shoal or nearest neighbour in the left arm is quicker (figures 2F-G). Including the speed of the focal fish to both models was believed to improve results, as quicker fish may be more likely to follow quicker groups. However, the results are once again poor, as the newly added variable caused a decline in AICc score from 1953 to 1951 in the NNS model, and from 1956.5 to 1951 in the ANS model. Furthermore, there was no statistical significance in both models for the focal speed variable (p $>$ 0.05, table 1).

In the ANO model, the focal fish was more likely to follow the shoal with higher polarization; thus more attracted to shoals swimming in a similar direction to them. The opposite therefore happened when a shoal was swimming in an opposing direction, as a repulsion seems to form. We also notice that this repulsion is stronger than the attraction mechanism discussed, as the slope of both curves favours the smaller angles (figure 2H). However, when looking at the NNO model, the results somewhat disagree with what has been established in the ANO model. This is because the values for the total fish swimming in the opposite direction to the focal fish do not affect the probabilities very strongly (pright = 0.083, pleft = 0.044), while the values for the fish swimming in the same direction do (pright & pleft $<$ .001). The coefficients for the "away" variables (same direction) in the model translate this stronger effect (table 1) and the logistic curves show the weaker predictions established with the "towards" variables (opposite direction) (figure 2I).

3.3 Comparison of Models

AICc scores for each model were calculated and compared to a base model containing the added effects of sex, trial and river, but without any explanatory variables related to social information (Decision $\sim$ 1). Figure 3 is a comprehensive review of these scores. The only model with an AICc lower than that of the base model was the average neighbour speed model (ANS), while the neighbour speed (NNS) performed slightly better. The best performing was the NND model, with an AICc difference of 300.9 from the base model, with AND being a close second. Unexpectedly, the ND model scored much lower than the two distance-related models. Another surprising result was the NN model, which scored very low. Finally, the novel orientation-based models show appropriate results to what was expected from the initial analysis, with much lower scores to the base model, but not as good as the metric distance-based models of NND and AND.

3.4 Combining Predictors

The data was fit to one last model with the better performing predictors for stronger predictive results. In order to avoid overfitting and multicollinearity, the variables were simplified down by subtracting the right arm data to the left arm data (where appropriate). Due to the similarities between the two orientation models, only the better performing model (NNO) was considered. NNS was not included due to the poor results in previous analysis, while ANS was included due to a possible link with average distance, and due to an decrease in AICc score by 12. Multivariate logistic regressions assume no multicollinearity in the variables. Correlation tests were therefore carried out between all the predictors using Variance Inflation Factors (VIFs). The count model had strong similarities to the “orient towards” variable and the “orient away” variable (figure 4A). Due to the latter having the most statistical significance in the previous fitted model, counts and “orient towards” were also not included. This final model performed very well, scoring 108.5 AICc less than the previous best model (NND). All variables had VIF values under 5, confirming no multicollinearity in the explanatory data (figure 4B). In our model comparison, we found ANS to be worse than our base model. However, when included with all other predictors, the average neighbour speed shows statistical significance in predicting fish decision-making (p = 0.002). All other variables improve the predictions as well, as we expected from the previous analysis, and their p-values confirm their statistical significance (see table 2). The probabilities of turning left in events identical to those in our dataset were predicted using our final model, and then compared to the actual decisions taken by the focal fish in those events. These values visualized suggest strong model predictions (figure 5). Furthermore, a Hosmer-Lemeshow test was performed to determine the goodness-of-fit of this model, and the results show no evidence to suggest a poor fit (χ^2 = 9.78, p = 0.281).

3.5 ML Classification

We have established logistic regression models which provide evidence for the importance of social information in the decision-making of guppies. However, this model provides probabilities for the likelihood of turning right or left. As an extension to this model, and in order to investigate the likelihood of randomness in turning decisions, the right and left arm labelled data was fitted to supervised learning classifiers as a model-free approach to predicting turning decisions. The variables used were the same as those in the final logistic regression model, therefore simplified down, in order to create a classification model with the variables being mostly independent from each other. The three algorithms of interest were: the K-Nearest Neighbour algorithm, due to its non-parametric nature (thus making no assumptions on the dataset), strength in smaller datasets such as this one ($<$ 100,000 labelled data samples), and its previous uses in similar studies [30]; Decision Tree Classifier, due to its uses with categorical data in regression or classification; and finally, the Naïve Bayes classifier, due to its uses with normally distributed predictor data (Gaussian NB) and, again, its functions with categorical data. Prior to any scaling, the KNN classifier scored 0.672, the Decision Tree classifier scored 0.652, and the Gaussian NB classifier scored 0.702. The most accurate results with differing groups of training data were therefore those gathered from the Gaussian Naïve Bayes classifier, most probably due to the normally distributed explanatory data. A principal component analysis concluded that there was no need to summarize the information content to a smaller dataset, and that 5 PCA components (therefore the same as the number of variables in the model) produced the best predictions. Using a standard scaler, various trials using different sets of training and testing data were conducted to determine the accuracy of the classifier. In most trials, the score ranged from 0.7 to 0.8. Figure 6 depicts the total predictions of the classifier in 50 separate trials, which show correct predictions in 75.5% of decision-making events (a 5.3% improvement from the classifier with non-scaled data). These results imply that about one in four turning events could not be predicted from the dataset, meaning some factors that are not possible to calculate through tracking data, or the individual differences between fish in a shoal, are probably also affecting individual decision-making.

4 DISCUSSION

The extent to which animals use social information to make decisions is key in studying collective animal behaviour [36]. This study demonstrates the influence of social information in the decision-making process of fish. No bias towards either directional option was found in the absence of social information. Furthermore, all the models in this study show no bias towards either decision when social information from the two options is identical, while also showing how differing social information can heavily influence decision-making. Due to the non-manipulative nature of this study, these results provide insight into the complex interaction rules that fish use in social contexts without any external stimuli [37]. The complicated mechanisms of animal living make it impossible for statistical modelling to accurately describe all these social interaction rules. However, due to the binary nature of the given options for each fish, the use of logistic regressions was possible, therefore providing probabilities for the decision-making of fish in simple decision-making events rather than definitive predictions which can often be misleading.

4.1 Metric ranges

Spatial positioning is an important factor in the social dynamics of any living organism, including fish [24–26, 38]. Here, metric range models were used to test the importance of this factor in guppy decision-making. As stand-alone factors, the nearest neighbour distance and the average neighbour distance models scored the highest in our AICc comparison. The NND model in particular showed very strong results, with some cases predicting that a fish will turn left up to 98% of the time if the NND in the left arm is much lower than that in the right arm. These results are not a surprise, as fish are capable of accurately estimating distances to their neighbours [19]. It therefore seems natural to assume that nearby fish will have a larger impact in decision-making. In a more adaptive perspective, we expect the distance of a fish to the rest of a shoal to be an important factor in survival, as a more cohesive shoal will perform better in predator evasion and foraging [3, 4]. Furthermore, metric range models state that animals follow all neighbours within a fixed distance [17], which is somewhat in line with the AND model. Therefore, as expected, this model had very strong results as well, suggesting that fish tend to follow the group that on average is closer to them. However, when discussing the attraction and repulsion zones of metric ranges [29], we can see from these results that further distances do not attract fish. We can therefore conclude that, in cases of binary decision-making, these zones of attraction are much smaller than anticipated, while there was no evidence for repulsion at any distance; instead, far away fish were more so ignored.

4.2 Quorum Responses & Last Observed Choice

As an alternative to metric ranges, studies have discussed counting systems (based on the ability of fish to differentiate between group sizes) to have an impact on decision-making. This theory has introduced topological ranges, where fish are believed to follow a set number of neighbours at any distance [13], and quorum responses, where individuals respond only when they see a threshold number of individuals perform a particular behaviour [20–22]. In this study, count data for either arm was fit to a logistic regression model to account for theories such as these, which will test whether fish are more likely to follow bigger groups. In cases of continuous decisions, therefore when all the fish were present in one arm, thus ignoring the other, focal fish were more likely to keep up with the cohesive shoal as the number of fish in this option increased. However, as soon as at least one fish was present in the other option, therefore making this a binary decision, the difference in counts was not as likely to affect decision-making. This suggests that fish may use counting systems in self-organizing contexts to maintain cohesion, as suggested by quorum decisions, but may not use these systems in binary decisions, therefore not following the assumptions of previous models suggesting that fish decisions can be explained using the differences of the number of animals choosing each option [10, 39, 40].

A simpler rule, “copy the last observed choice”, has also been suggested and studies have found it to give similar predictions to those of other established optimal models [23]. Here, the last choice of the neighbour leading in front of the focal fish was fit to a logistic regression as a binary variable (left arm or right arm). The predictions suggest that fish copy this last observed choice about 60% of the time in binary decisions, but this information is not as useful in decision-making as others we have discussed. In continuous decisions (all fish in one arm) these predictions were not affected, suggesting that copying the last observed choice is not a better determinant in maintain cohesion rather than decision-making. While this model meets the basic assumptions of topological ranges, which assume that animals will follow N number of nearest neighbours [13], a much larger shoal would open the possibility to more accurate testing of topological ranges by testing larger fixed numbers of neighbours.

4.3 Speed & Orientation

We discuss speed as a novel interaction rule in decision-making, under the assumptions that fish have a preference for quicker shoals due to the higher likelihood of nearby food sources, as well as healthier individuals [4, 31]. Results did suggest a slight preference for quicker shoals, as expected. However, the statistical significance of this preference was very low. As a stand-alone determinant, speed is therefore unlikely to influence fish decisions. We there-

fore introduce a second novel interaction rule for decision-making, based on the orientation of nearby neighbours. Orientation has widely been discussed as a means of parallel alignment in neutral zones for better cohesion [24, 25]. This study instead focuses on the uses of orientation as a stand-alone useful piece of social information in decision-making, as we believe fish tend to follow the shoal which more closely matches their swimming direction. The first model provides strong evidence for attraction and repulsion between fish according to their swimming directions (aligned or misaligned directions, respectively). The second model, while still providing some evidence for repulsion mechanisms, mostly suggests that fish ignore their neighbours swimming in the opposite direction, while instead following the group that had more fish in the same direction. These final results suggest the use of quorum responses in decision-making, when taking into account that some fish will be ignored due to their orientation. In other words, fish will make decisions based on the number of fish that are swimming in their same direction and have already taken that decision.

4.4 Final Models

Following thorough model comparisons, as well as collinearity analysis to ensure independent determinants for decision-making, we determined that the best model using the information extracted from tracking data included the ND predictor (copy the last observed choice); the NND and AND predictors (follow nearby fish); the NNO (away) predictor (follow bigger group of fish oriented away); and, unexpectedly, the ANS predictor (follow quicker group). An interaction between the ANS and the AND predictors proved to be significant in the model, suggesting that fish do in fact follow quicker groups, probably due to an assumed nearby food source, granted the group is near enough for a focal fish to notice this difference in speed. As a confirmation of these final predictors, a Gaussian Naïve Bayes Classifier was fit with this data and the results are as expected: in about three out of four decisions, the classifier could accurately predict the direction of a fish making a decision. The incorrect predictions can be said to be due to unavailable data, such as individual differences in the fish, or other social cues that could not be collected from the available data.

4.5 Conclusion

The study of social animal decision-making has functions relating to human engineering through the philosophy of biomimetics, which is the synthesis of materials and machines that mimic biological processes [41]. This study has confirmed the nearest and average neighbour distances as crucial determinants of decision making. It has also introduced speed as a useful determinant in some cases of decision-making, while discussing the possible importance of orientation in creating attraction and repulsion zones. More importantly, orientation-based quorum re-

sponses have been introduced as a novel determinant in decision-making, with very promising results to be further investigated in the future. While the use of binary decision-making events made the comparison of various determinants possible, it is limiting in studying other determinants such as topological ranges, which require larger groups. Using a different approach to determining these social cues would therefore probably lead to somewhat different conclusions. For example, in cases of more complex decisions, these determinants may not be as significant. A five-arm water maze may be of interest to determine the differences in decision-making interaction rules in more complex scenarios, while maintaining a similar approach to this study.

REFERENCES

[1] I. Giardina, “Collective behavior in animal groups: Theoretical models and empirical studies,” *HFSP Journal*, vol. 2, no. 4, 2008.

[2] J. K. Parrish and L. Edelstein-Keshet, “Complexity, pattern, and evolutionary trade-offs in animal aggregation,” *Science*, vol. 284, no. 5411, 1999.

[3] C. C. Ioannou, I. W. Ramnarine, and C. J. Torney, “High-predation habitats affect the social dynamics of collective exploration in a shoaling fish,” *Science Advances*, vol. 3, no. 5, 2017.

[4] R. Harpaz and E. Schneidman, “Social interactions drive efficient foraging and income equality in groups of fish,” *eLife*, vol. 9, 2020.

[5] J. P. Hollins, D. Thambithurai, T. E. Van Leeuwen, B. Allan, B. Koeck, D. Bailey, and S. S. Killen, “Shoal familiarity modulates effects of individual metabolism on vulnerability to capture by trawling,” *Conservation Physiology*, vol. 7, no. 1, 2019.

[6] M. Wolf and J. Krause, “Why personality differences matter for social functioning and social structure,” *Trends in Ecology and Evolution*, vol. 29, no. 6, 2014.

[7] G. Polverino, N. Abaid, V. Kopman, S. MacRì, and M. Porfiri, “Zebrafish response to robotic fish: Preference experiments on isolated individuals and small shoals,” *Bioinspiration and Biomimetics*, vol. 7, no. 3, 2012.

[8] J. D. Davidson, M. M. Sosna, C. R. Twomey, V. H. Sridhar, S. P. Leblanc, and I. D. Couzin, “Collective detection based on visual information in animal groups,” *Journal of the Royal Society Interface*, vol. 18, no. 180, 2021.

[9] L. Jiang, L. Giuggioli, A. Perna, R. Escobedo, V. Lecheval, C. Sire, Z. Han, and G. Theraulaz, “Identifying influential neighbors in animal flocking,” *PLoS Computational Biology*, vol. 13, no. 11, 2017.

[10] A. J. Ward, J. E. Herbert-Read, D. J. Sumpter, and J. Krause, “Fast and accurate decisions through collective vigilance in fish shoals,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 6, 2011.

[11] T. J. Pitcher, “Heuristic definitions of fish shoaling behaviour,” *Animal Behaviour*, vol. 31, no. 2, 1983.

[12] S. Gueron, S. A. Levin, and D. I. Rubenstein, “The dynamics of herds: From individuals to aggregations,” *Journal of Theoretical Biology*, vol. 182, no. 1, 1996.

[13] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic, "Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 4, 2008.

[14] J. L. Harcourt, T. Z. Ang, G. Sweetman, R. A. Johnstone, and A. Manica, "Social feedback and the emergence of leaders and followers," *Current Biology*, vol. 19, no. 3, 2009.

[15] Y. Inada and K. Kawachi, "Order and flexibility in the motion of fish schools," *Journal of Theoretical Biology*, vol. 214, no. 3, 2002.

[16] H. Kunz and C. K. Hemelrijk, "Artificial fish schools: collective effects of school size, body size, and body form," *Artificial Life*, vol. 9, no. 3, 2003.

[17] I. D. Couzin and J. Krause, "Self-organization and collective behavior in vertebrates," *Advances in the Study of Behavior*, vol. 32, 2003.

[18] J. H. Tien, S. A. Levin, and D. I. Rubenstein, "Dynamics of fish shoals: Identifying key decision rules," *Evolutionary Ecology Research*, vol. 6, no. 4, 2004.

[19] M. A. Goodale, C. G. Ellard, and L. Booth, "The role of image size and retinal motion in the computation of absolute distance by the Mongolian gerbil (Meriones unguiculatus)," *Vision Research*, vol. 30, no. 3, 1990.

[20] A. J. Ward, J. Krause, and D. J. Sumpter, "Quorum decision-making in foraging fish shoals," *PLoS ONE*, vol. 7, no. 3, 2012.

[21] D. J. Sumpter and S. C. Pratt, "Quorum responses and consensus decision making," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1518, 2009.

[22] A. J. Ward, D. J. Sumpter, I. D. Couzin, P. J. Hart, and J. Krause, "Quorum decision-making facilitates information transfer in fish shoals," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 19, 2008.

[23] K. Kadak and N. Miller, "Follow the straggler: zebrafish use a simple heuristic for collective decision-making: heuristics for collective choice," *Proceedings of the Royal Society B: Biological Sciences*, vol. 287, no. 1940, 2020.

[24] A. Huth and C. Wissel, "The simulation of the movement of fish schools," *Journal of Theoretical Biology*, vol. 156, no. 3, 1992.

[25] I. Aoki, “A simulation study on the schooling mechanism in fish,” *Nippon Suisan Gakkaishi*, vol. 48, no. 8, 1982.

[26] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks, “Collective memory and spatial sorting in animal groups,” *Journal of Theoretical Biology*, vol. 218, no. 1, 2002.

[27] V. Lecheval, L. Jiang, P. Tichit, C. Sire, C. Hemelrijk, and G. Theraulaz, “Domino-like propagation of collective U-turns in fish schools,” *bioRxiv*, 2017.

[28] Y. Katz, K. Tunstrøm, C. C. Ioannou, C. Huepe, and I. D. Couzin, “Inferring the structure and dynamics of interactions in schooling fish,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 46, 2011.

[29] G. Li, I. Ashraf, B. François, D. Kolomenskiy, F. Lechenault, R. Godoy-Diana, and B. Thiria, “Burst-and-coast swimmers optimize gait by adapting unique intrinsic cycle,” *Communications Biology*, vol. 4, no. 1, 2021.

[30] J. Gautrais, F. Ginelli, R. Fournier, S. Blanco, M. Soria, H. Chaté, and G. Theraulaz, “Deciphering interactions in moving animal groups,” *PLoS Computational Biology*, vol. 8, no. 9, 2012.

[31] C. Cano-Barbacil, J. Radinger, M. Argudo, F. Rubio-Gracia, A. Vila-Gispert, and E. García-Berthou, “Key factors explaining critical swimming speed in freshwater fish: a review and statistical analysis for Iberian species,” *Scientific Reports*, vol. 10, no. 1, 2020.

[32] S. W. Griffiths and A. E. Magurran, “Sex and schooling behaviour in the Trinidadian guppy,” *Animal Behaviour*, vol. 56, no. 3, 1998.

[33] A. Strandburg-Peshkin, C. R. Twomey, N. W. Bode, A. B. Kao, Y. Katz, C. C. Ioannou, S. B. Rosenthal, C. J. Torney, H. S. Wu, S. A. Levin, and I. D. Couzin, “Visual sensory networks and effective information transfer in animal groups,” *Current Biology*, vol. 23, no. 17, 2013.

[34] L. Lei, R. Escobedo, C. Sire, and G. Theraulaz, “Computational and robotic modeling reveal parsimonious combinations of interactions between individuals in schooling fish,” *PLoS Computational Biology*, vol. 16, no. 3, 2020.

[35] M. E. M. Petrazzini, V. A. Sovrano, G. Vallortigara, and A. Messina, “Brain and behavioral asymmetry: a lesson from fish,” 2020.

[36] J. Duboscq, C. Neumann, M. Agil, D. Perwitasari-Farajallah, B. Thierry, and A. Engelhardt, “Degrees of freedom in social bonds of crested macaque females,” *Animal Behaviour*, vol. 123, 2017.

[37] J. Krause, G. D. Ruxton, and S. Krause, “Swarm intelligence in animals and humans,” *Trends in Ecology and Evolution*, vol. 25, no. 1, 2010.

[38] C. W. Reynolds, “Flocks, herds and schools: a distributed behavioural model.,” *Computer Graphics (ACM)*, vol. 21, no. 4, 1987.

[39] S. Arganda, A. Pérez-Escudero, and G. G. De Polavieja, “A common rule for decision making in animal collectives across species,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 50, 2012.

[40] A. Pérez-Escudero and G. G. de Polavieja, “Collective animal behavior from bayesian estimation and probability matching,” *PLoS Computational Biology*, vol. 7, no. 11, 2011.

[41] J. Hwang, Y. Jeong, J. M. Park, K. H. Lee, J. W. Hong, and J. Choi, “Biomimetics: forecasting the future of science, engineering, and medicine,” *International Journal of Nanomedicine*, vol. 10, 2015.

5 TABLES & FIGURES

Model	Predictors	Intercept	Coefficients	Coefficients p-value	AICc	Odds Ratio (OR)	95% CI of OR
NN (DMEs)	Right NN	0.022 ± 0.076	-0.276 ± 0.021	<.001	4573.9	0.759	(0.791, 0.728)
	Left NN		0.253 ± 0.021	<.001		1.288	(1.342, 1.236)
NN (bDMEs)	Right NN	0.022 ± 0.195	-0.183 ± 0.043	<.001	1916.3	0.833	(0.906, 0.765)
	Left NN		0.125 ± 0.438	0.0045		1.133	(2.674, 0.48)
NND	Right NND	-0.173 ± 0.143	0.861 ± 0.072	<.001	1630.8	2.366	(2.724, 2.054)
	Left NND		-0.879 ± 0.073	<.001		0.415	(0.479, 0.36)
ND	ND	-0.612 ± 0.104	0.953 ± 0.111	<.001	1881	2.593	(3.224, 2.086)
AND	Right AND	0.028 ± 0.168	0.653 ± 0.066	<.001	1729.6	1.921	(2.187, 1.688)
	Left AND		-0.742 ± 0.067	<.001		0.476	(0.543, 0.418)
NNS	Right NNS	0.003 ± 0.142	-0.270 ± 0.117	0.021	1953	0.763	(0.96, 0.607)
	Left NNS		0.133 ± 0.116	0.252		1.142	(1.434, 0.91)
ANS	Right ANS	0.026 ± 0.154	-0.245 ± 0.146	0.093	1956.5	0.783	(1.042, 0.588)
	Left ANS		0.068 ± 0.157	0.663		1.07	(1.456, 0.787)
ANO	Right ANO	-0.191 ± 0.183	-0.006 ± 0.001	<.001	1876.3	0.994	(0.996, 0.992)
	Left ANO		0.007 ± 0.001	<.001		1.007	(1.009, 1.005)
NNO	Right NNO (away)	0.184 ±0.189	-0.355 ± 0.052	<.001	1827.8	0.701	(0.776, 0.633)
	Left NNO (away)		0.269 ± 0.052	<.001		1.309	(1.449, 1.182)
	Right NNO (towards)		0.101 ± 0.058	0.083		1.106	(1.239, 0.987)
	Left NNO (towards)		-0.115 ± 0.057	0.044		0.891	(0.997, 0.797)

Table 1: Results for all logistic regression models. Coefficients, p-value of coefficients, odds ratio (OR) and 95% confidence interval (CI) of OR for all predictors in each model, along with intercept and AICc for each model.

Predictors	Intercept	Coefficients	Coefficients p-value	AICc	Odds Ratio (OR)	95% CI of OR
NNO(away)	-0.009 ± 0.125	0.346 ± 0.039	<.001	1522.3	1.413	(1.526, 1.309)
ANS		-0.527 ± 0.146	.002		0.59	(0.786, 0.443)
NND		-0.567 ± 0.095	<.001		0.567	(0.683, 0.471)
AND		-0.501 ± 0.082	<.001		0.606	(0.712, 0.516)
ND		-0.484 ± 0.159	<.001		0.616	(0.842, 0.451)

Table 2: Results for the final combined logistic regression model, including the coefficients, the p-value of the coefficients, the OR and the 95% CI of the OR for each predictor, along with the intercept and the AICc for the model. The predictors are the difference of the right arm values from the left arm values of NNO (away), ANS, NND, AND and ND.

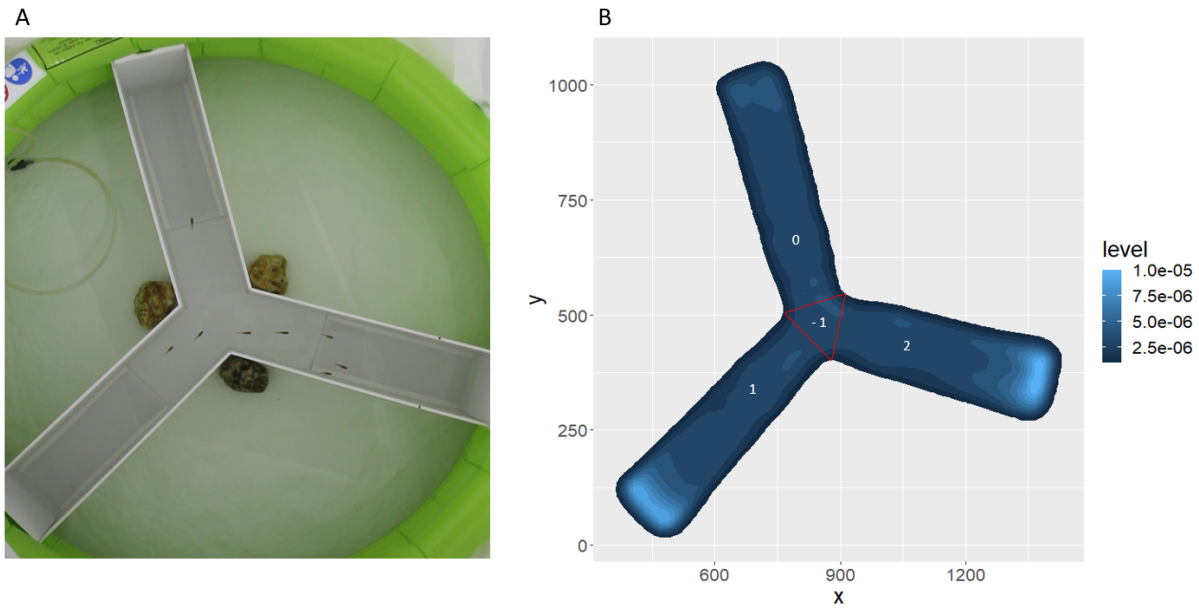

Figure 1: *Y-maze used for collection of tracking data. A) One frame of y-maze from the top during a decision event (red circle indicates fish initiating a decision-event) with neighbours present in both the right and the left arm. B) Heatmap of all positions of fish in all 42 trials, with more fish clearly present at the ends of the arms as seen from the brighter blue (due to the shoal slowing down to turn around).*

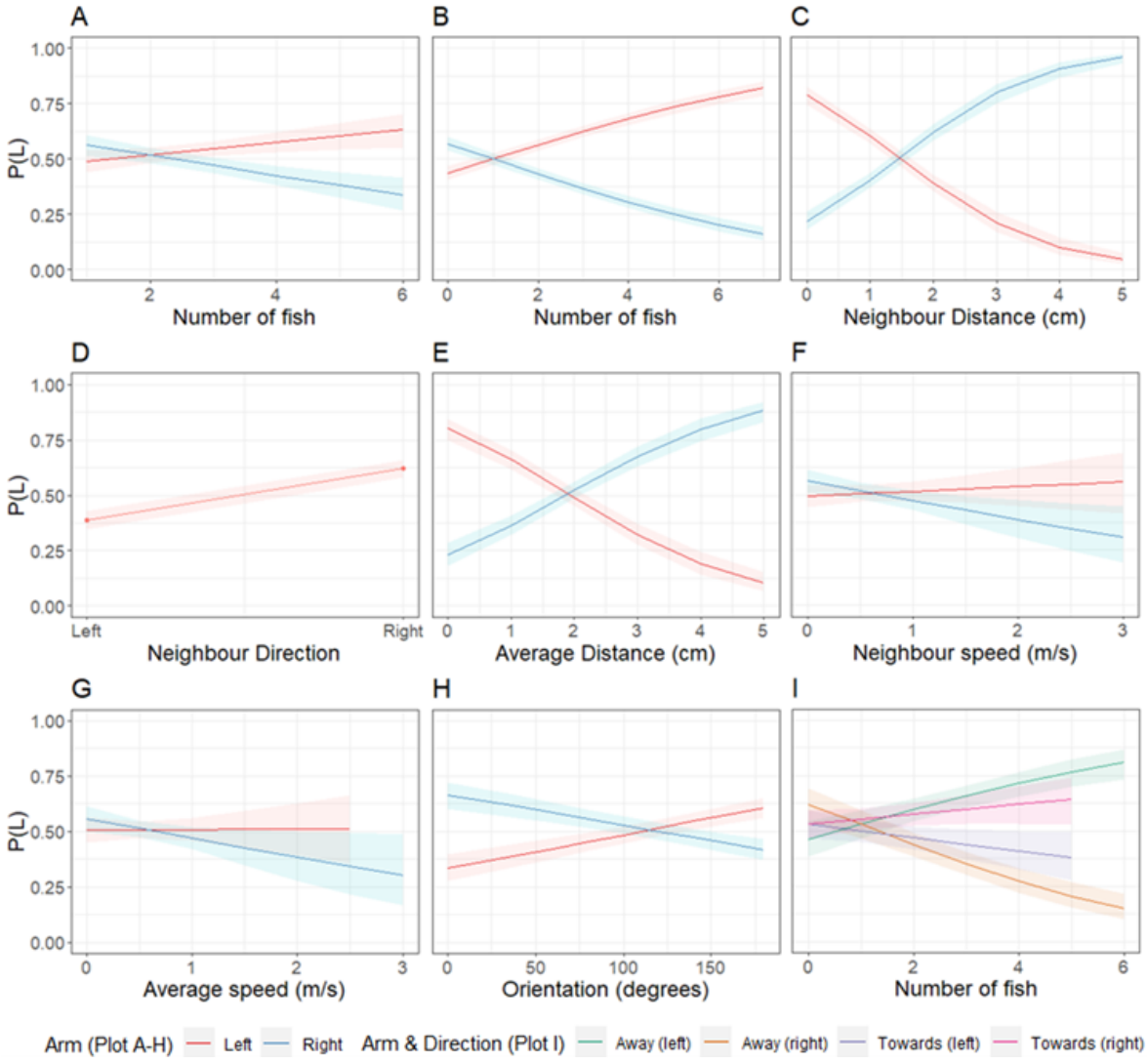

Figure 2: *Predicted probabilities P(L) of turning left for every model (right arm values vs left arm values). NN model for all decisions (A) and for binary decisions only (B): models based on counting systems in fish. C) NND model; based on metric ranges (nearest neighbour). D) ND model; based on "copy the last observed choice". E) AND model; based on metric ranges (all neighbours). F) NNS model; based on speed of nearest neighbour. G) ANS model; based on average speed of all neighbours. H) ANO model; based on orientation (using degrees). I) NNO model; based on orientation (using fish counts).*

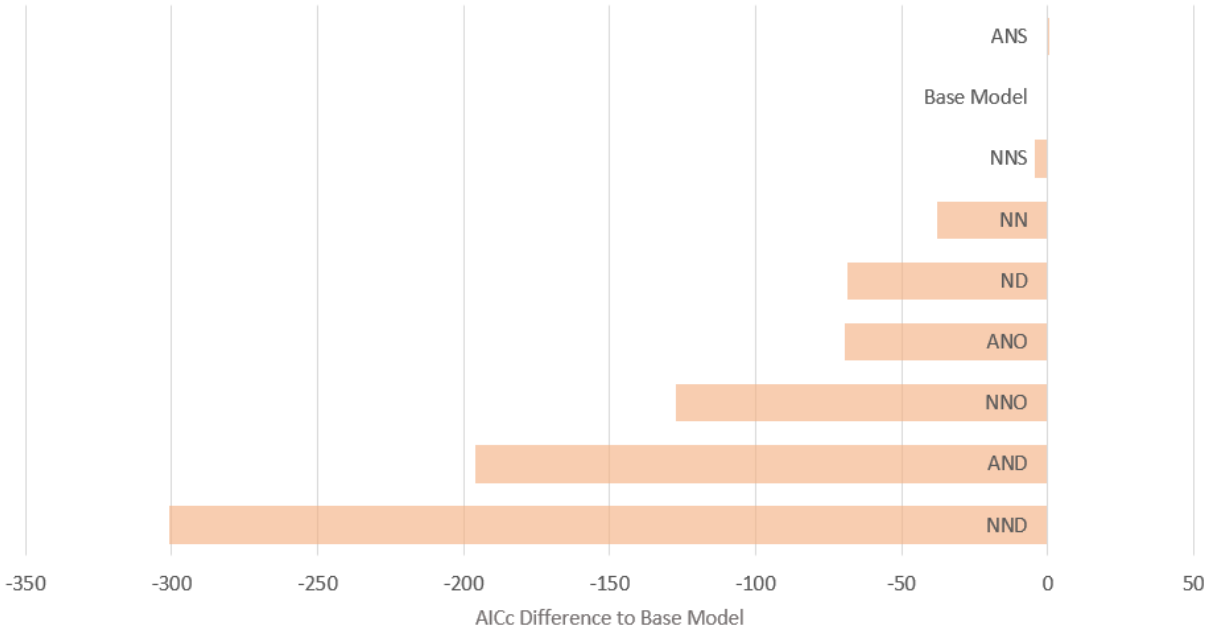

Figure 3: *Comparison of AICc model score from base model (Decision ∼ 1). Each AICc was subtracted from the base model AICc. The bigger the negative value, the better the model score. A positive AICc difference indicates a model that performs worse than predicting values through randomness.*

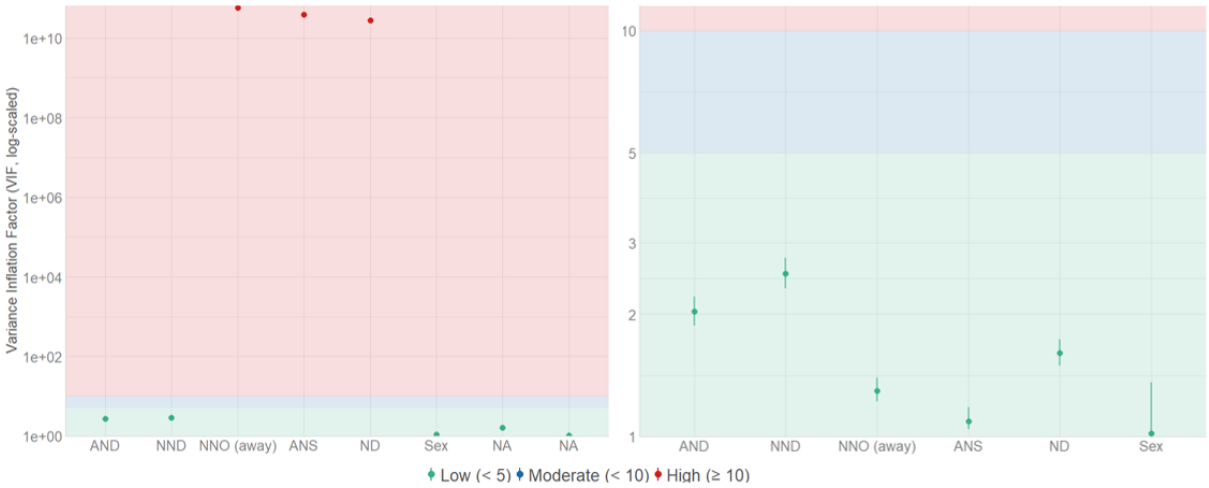

Figure 4: *Predictor VIF scores for two models. In the first (A), NNO(away), ANS and ANO have very high VIF scores ($>$ 10). The second model (B) therefore does not include the ANS and ANO predictors due to their strong collinearity to NNO (away). The second model shows low VIF scores suggesting no collinearity.*

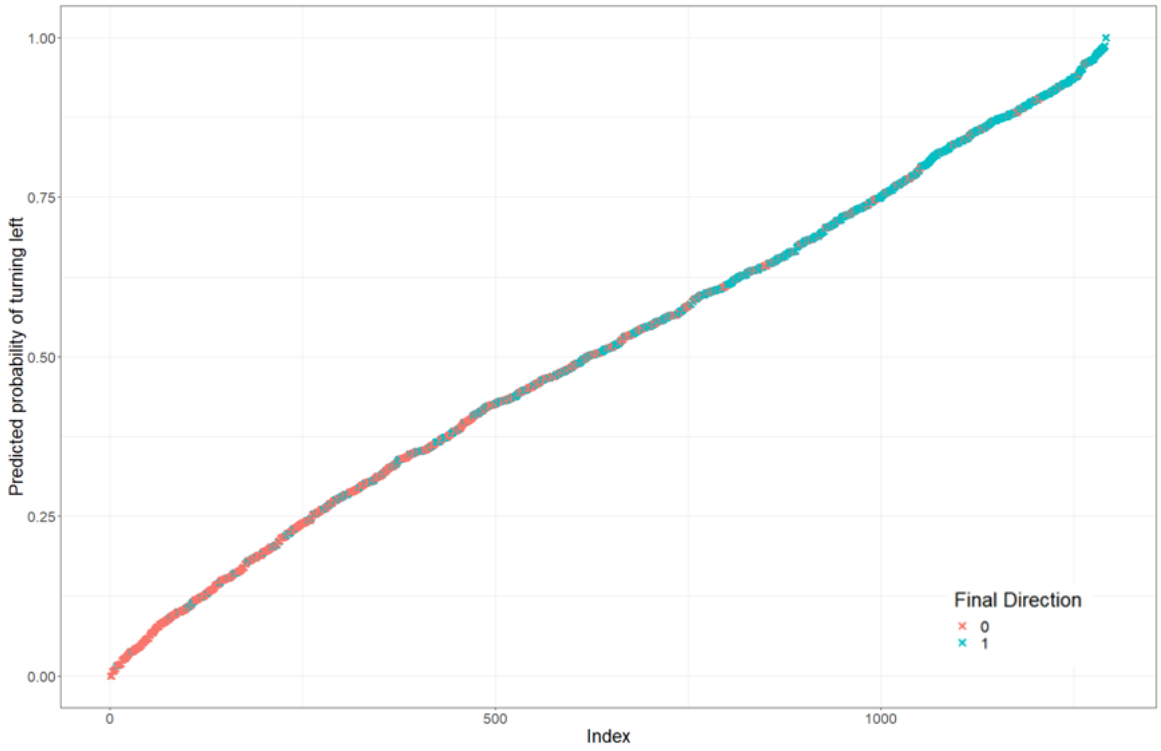

Figure 5: *Logistic curve of the predicted probabilities of turning left in the combined model. Values from actual decision-making events were used to predict the probability of turning left. These are then compared to the real direction of the focal fish in these decisions (red for right, blue for left). When P(L) $<$ 0.5, it is more likely fish will turn right (final direction = 0); when P(L) $>$ 0.05, it is more likely fish will turn left (final direction = 1).*

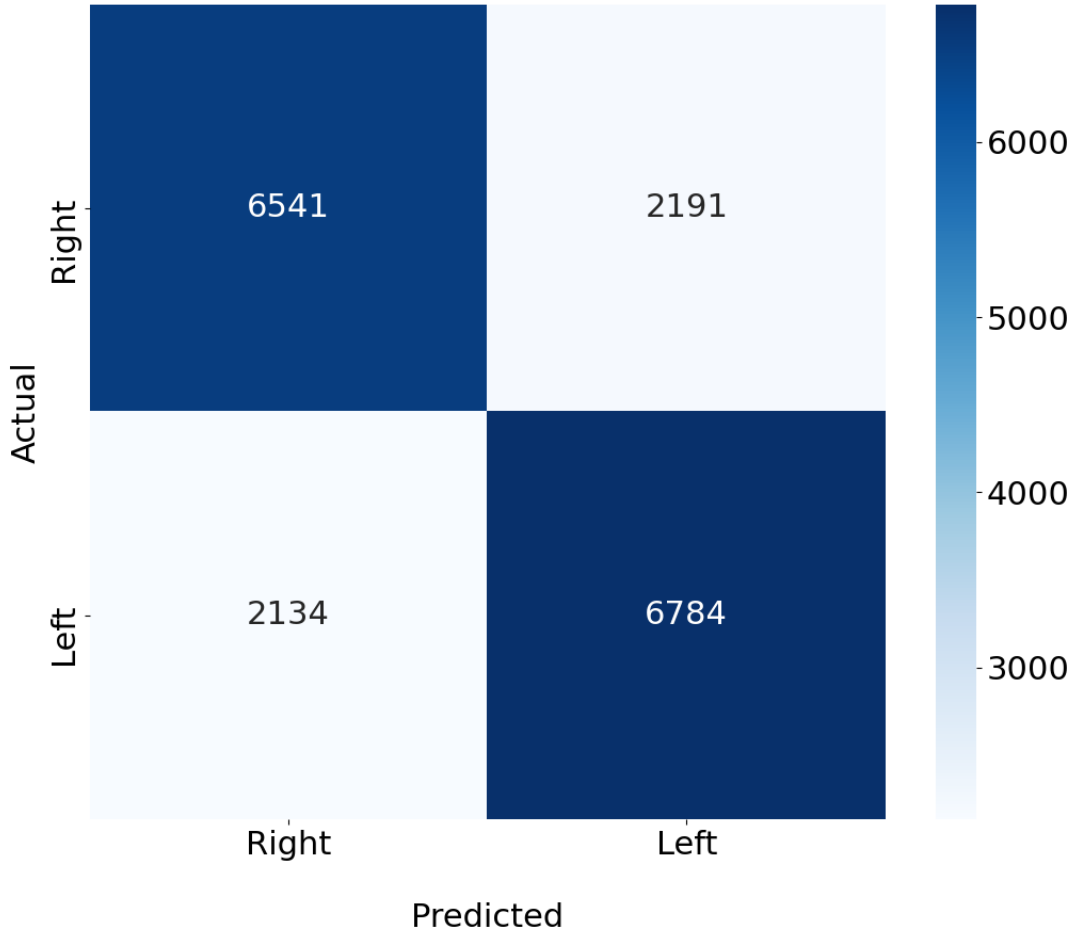

Figure 6: *Confusion Matrix depicting the predicted and actual values for a scaled Gaussian Naive Bayes classifier (PCA = 5). The classifier predicts the binary turning decisions of eight fish in a water maze. The confusion matrix shows the predicted values after 50 trials of the classifier with different sets of training and testing data (total predicted decisions after 50 trials = 17,650).*