Bayesball: Bayesian Integration in Professional Baseball Batters

Justin A. Brantley [1,✉] **and Konrad P. Körding** [1]

[1]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104

Pitchers in baseball throw the ball with such high velocity and varying movement that batters only have a few hundred milliseconds to estimate whether to swing and how high to swing–contacting the ball too high or too low may produce hit balls that easily result in an out. Even before the pitcher releases the ball, the batter has some belief, or estimated distribution (a 'prior'), of where the ball may land in the zone. Batters will update this prior belief with information from observing the pitch (the 'likelihood') to calculate their final estimate (the 'posterior'). These models of behavior, called Bayesian models within movement science, predict that when players have better prior information, e.g. because they know the upcoming pitch due to 'tipping', that they will rely more on prior information; by contrast if their prior is less informative, e.g. because the pitch is very random as in the case of a knuckleball, they will instead rely more on the observation. Here we test these models using information from more than a million pitches from professional baseball. We find that batters integrate prior information with noisy observations to manage pitch uncertainty. Moreover, as predicted by a Bayesian model, a batter's estimate of where to swing is biased towards the prior when the pitch is tipped and biased towards the likelihood in the case of pitches with high uncertainty. These results demonstrate that Bayesian ideas are relevant well beyond laboratory experiments and matter in the world of sports.

Bayesian brain | baseball | movement
Correspondence: ***justin.a.brantley@gmail.com***

Introduction

> *You can't think and hit the ball at the same time*
>
> **—Yogi Berra**

It is remarkable that baseball batters can successfully hit the ball so often despite the uncertainty in pitched balls. With many pitchers averaging pitch velocities of over 95–100mph[1] thrown over a distance of 60.5 feet (Figure 1A), professional batters have approximately 400 milliseconds to perceive the ball, decide on an action, and swing[2]. Their perceptual system is limited, thus the ball can only be observed for a short period of time until the brain needs to make a decision. It takes approximately 100-ms to observe the ball and an additional 150-ms to physically swing the bat over the plate[2]. This leaves 150–250-ms between observation and movement initiation to decide whether or not to swing. In such cases, there will always be uncertainty about the specific trajectory of the pitched ball. The pitcher is trying to maximize this uncertainty by spinning the ball in various ways to introduce curved trajectories that are hard to predict. They purposefully introduce multiple sources of uncertainty (e.g., pitch selection, velocity, and spin rate—Figure 1A-B) through deception and physical perturbations on the thrown ball. Pitchers try to make distinguishing between trajectories hard by moving their bodies in similar ways across wildly different pitch types. At the moment when the ball leaves the pitcher's hand, the batter can at best estimate a probability distribution of where the ball may end up. This distribution is called the prior (Figure 1A, right, shown in blue) and is one of the pieces of the puzzle of estimating where the ball will land.

The second part of the puzzle of where the ball will land is the observation of the trajectory of the ball. Based on observing the ball after the pitch is thrown, the brain could form a likelihood distribution (Figure 1A, right, shown in grey) representing how probable the observations of the pitched balls are under which assumption of the thrown ball trajectory (ignoring the prior). This information will also not be perfect, after all, our eyes have limited resolution and our brain has limited processing speed relative to the high velocity of the pitch. However, information from seeing the ball, such as release point, the pitcher's grip on the ball, spin direction, and ball speed, is clearly useful and may be summarized by the so-called likelihood function (Figure 1A, right, shown in grey). This defines how compatible each observed ball trajectory is with a potential height within the strike zone. Despite all of these uncertainties, the batter usually makes reasonable estimations of the ball's trajectory and are still able to achieve hits approximately 25% of the time on average across all batters, with some exceptional players reaching rates of over 30%. How can a batter predict a ball's trajectory so well that they hit the ball successfully? More generally, we may ask how baseball batters estimate the pitched ball's trajectory if all they have is noisy prior information and the noisy sensory information from watching the pitched ball?

Bayesian statistics is the discipline that formalizes how the two sources of information, the prior knowledge and observations of the thrown ball, must be combined. It

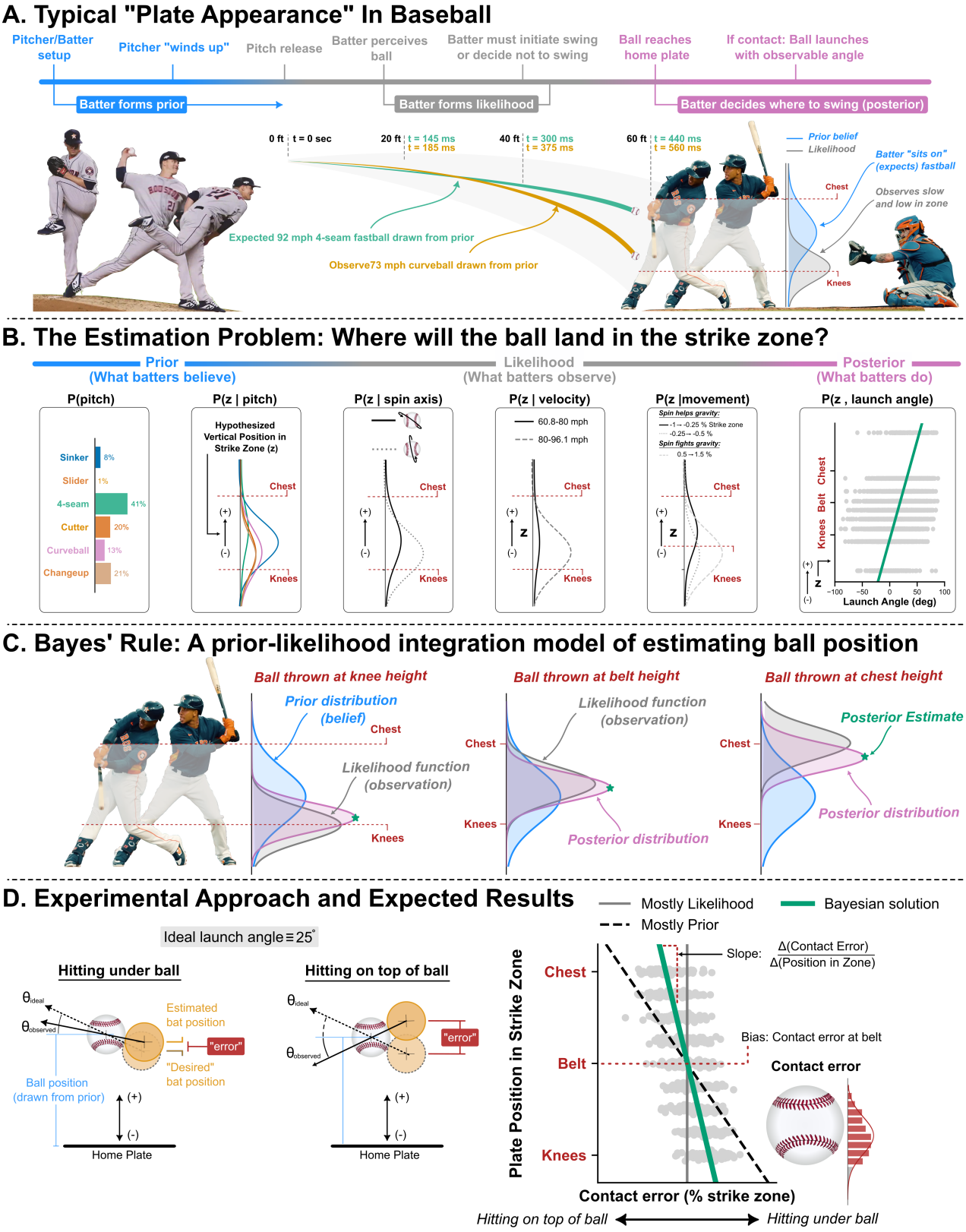

Figure 1. Testing Bayesian ideas by analyzing Baseball data. A. In a typical plate appearance, a batter may form a prior over where the ball will land in the strike zone. They will subsequently observe the ball's trajectory and form a likelihood distribution for the final location, i.e., the probability given the observation of the pitch. **B.** A batter's belief about where the ball will land is based on the pitcher's unique pitch behavior, such as pitch selection, velocity, spin rate, and movement. **C.** Bayes' rule allows for the combination of prior knowledge with observations to make an estimate, called the posterior. **D.** We experimentally test whether batters behave in a Bayesian way by estimating the contact error on batted balls. Our theoretical model illustrates a Bayesian solution, a mostly likelihood model, and a mostly prior model.

states that the prior distribution and the likelihood function must be multiplied to gain a posterior estimate of the ball trajectory[3–6] (Figure 1C). The result is a situation where what we know from observing the pitcher (prior) and what we observed during the ball's flight (likelihood) are combined based on the level of uncertainty in each. For example, when vision of the ball is very good (e.g. because the ball is slow or the spin pattern is clear), batters should rely mostly on how they see the ball fly. On the other hand, when the pitcher is more predictable, batters should rely more on their prior knowledge. In the language of Bayesian statistics, we might say that a batter combines their prior beliefs about a pitch, $\mathbf{P}(A)$, with the likelihood, or probability of the observed pitch given that the belief is true, $\mathbf{P}(B|A)$. They then make a prediction, $\mathbf{P}(A|B)$, which is the posterior probability of the pitch after taking into account the observation of the pitch. Formally, we express this as

$$\mathbf{P}(A|B) \propto \mathbf{P}(B|A) \cdot \mathbf{P}(A).$$

Thus, Bayesian statistics provides us with a theory of how the brain navigates decision making and movement in an uncertain world. In the case of batters, we theorize that Bayesian statistics provides a framework for understanding how batters hit pitched balls in the face of uncertainty. However, this begs the question: are humans and their behavior like Bayesian statistics? Several hypotheses may explain how batters deal with pitch uncertainty during batting. The first is by only considering prior information. By generating a prior distribution of pitches, the batter might then select some action that maximizes the probability of hitting under the prior while making no adjustments to the observations of the thrown pitch. Under this assumption the slope of the regression line in our analysis (and prior studies[4]) would increase towards one (dashed line in Figure 1D). A second approach is that batters could rely exclusively on observable data from each pitch without using any prior knowledge for their batting estimate (likelihood only–solid vertical line with zero slope in Figure 1D). While possible from a computational perspective, this hypothesis is unlikely to accurately describe a batter's behavior since it assumes that batters do not make estimates about the visual uncertainty based on prior information. According to this strategy, the error around the estimate would vary between pitches, but the average error would remain constant for all batting estimates (Figure 1D.). A third strategy—Bayesian statistics—predicts that a batter will combine prior information with the likelihood to make an inference about the oncoming pitch. This approach is favorable in that it allows for the use of prior information about pitch behavior while permitting flexibility in the estimate by accounting for uncertainty in both the prior and observation. Practically speaking, this is expected as we know that batters study opposing pitchers in advance of the game (prior information) and that batters often achieve optimal outcomes despite the varying sources of uncertainty on the pitch. Moreover, the physical distance of 60.5 feet between the pitcher and batter further evidences the need for Bayes-like solution to the estimation problem: since a ball thrown at 95-mph takes approximately 450-ms to reach home plate, a batter only has 150-250 ms to decide precisely where to swing to contact the ball. If instead, the pitcher were to throw the equivalent pitch from the middle of center field ($\approx$ 250 feet away), the batter would have more than four times the amount of time to observe the pitch and would thus have no use for a prior belief. In contrast, if the batter were to throw from half the distance that they normally do, the batter has barely enough time to physically swing the bat to the correct position in the zone, requiring them to make an estimate based on pure prior belief about the location of the pitch. To test Bayesian ideas we need to know about

(1) the existence of the prior distribution (e.g. from the distribution of heights of pitches within the strike zone);

(2) the results of a perturbation (the final vertical position of the ball); and, importantly,

(3) we need to know the error a subject makes when making an estimate.

The first two can be directly measured, however, here we make an approximation for the estimation error on batted balls. Based on recent trends in baseball batting mechanics, we assume that batters aim to contact the ball at the point on the ball that maximizes the probability of a home run[7]. We assume a simple bat/ball contact model based on the geometry of the ball and the bat and compute an error on the contact based on the actual launch angle and the "optimal" launch angle (Figure 1D). We can thus approximate an estimation error and have a way of testing predictions from the Bayesian framework. A Bayesian approach to sensorimotor control has been extensively documented through decades of laboratory studies on movement and decision making[4–6,8,9]. A common experimental approach to test this hypothesis for motor control is a 2-dimensional reaching task, e.g., using a robotic manipulandum, a tracking system, or a computer cursor to capture the movement behavior, where the subjects are asked to reach towards a target while a perturbation is applied to the movement path. The perturbations are drawn from some prior distribution and feedback with varying levels of uncertainty is provided to the subjects about their action. Since the correct path can only be noisily observed (through the feedback with varying uncertainty), the subject should theoretically alter their movement path to correctly execute the task based only on their prior information and the noisy feedback. Indeed, subjects behave in a way consistent with Bayes rule, where they rely more on prior knowledge when observing feedback with high uncertainty and rely more on the feedback when it contains lower uncertainty. While these studies suggest that the brain im-

plements some form of Bayesian integration for managing uncertainty, "Bayesian brain" behavior has scarcely been documented in real world data. Moreover in addition to observing this behavior, we ask, does Bayesian behavior matter for real world movements?

Here we use the batting behavior of professional baseball players to test if humans use Bayesian approaches in the real world. Specifically, we hypothesize that batters combine prior information about the pitch with observations during each time they bat to achieve hits, and that this behavior can be explained using a Bayesian framework. We use publicly available data recorded from professional Major League Baseball (MLB) games in the United States. The data were acquired using MLB's tracking technology, Statcast[10], which relies on high speed cameras and radars to record play-by-play actions. We find that the behavior observed in our baseball data is consistent with laboratory experiments and implies that batters combine prior information with noisy observations to manage pitch uncertainty in a way that is consistent with Bayesian statistics. Furthermore, when batters have prior information about the oncoming pitch, our results show that they rely more on the prior knowledge. Moreover, we show that for certain pitch types with high probability yet high behavioral uncertainty (that is, batters know what pitch is coming but its movement is noisy), batters rely nearly completely on observations about the oncoming pitch while ignoring the uninformative prior knowledge associated with the pitch type. These results demonstrate Bayesian behavior in a real world skilled movement task.

Results

Here we test if baseball batters employ a strategy for managing pitch uncertainty that can be explained by Bayesian statistics. We gathered data from MLB's publicly accessible data clearinghouse[11] for MLB regular and postseason games from 2015-2021. After cleaning the data to remove out-of-season games and plate apperances where no contact was made, we were left with 1,036,405 pitches for analysis. Within this data set we can distinguish different pitch types, giving us variables that may affect the prior uncertainty of pitches. With this, we can thus approach asking questions about uncertainty processing by professional batters. Estimating vertical pitch position is challenging for the batter because it varies from pitch to pitch. We directly have this information in our data. The overall distribution of the vertical pitch position was roughly normally distributed, $\mathcal{N}(\mu = 0.45, \sigma = 0.32)$, with approximately 88% located within the vertical limits of the strike zone (Figure 2A). This distribution of vertical pitch position is crucial as it defines the distribution of relevant variables to be estimated. Pitchers use the entire strike zone, and even beyond, to make the estimation task difficult for the batter. The basic prediction of Bayesian statistics is that the further up the ball is within the strike zone, the more strongly the batter should estimate a position biased downward and vice versa. We can test this idea by plotting the inferred error as a function of the position within the strike zone. We computed the estimate of the error as the difference between the "ideal" contact point on the ball (i.e., the point on the ball that results in a launch angle that maximizes the probability of a home run) and the observed contact point (Figure 1D). It is important to note that in our regression model, we treat the location within the strike zone as the explanatory variable and the error on the batted-ball as the response. However, the axes are reversed in the figures to improve visual clarity when considering location within the strike zone relative to the batter. Indeed, we find a systematic correlation between position and error as predicted by the theory (Figure 2B), suggesting a Bayesian strategy for integration where both prior and likelihood are used for estimation.

External cues impact the strength of the prior

Most major league pitchers throw an average of 4-5 pitches, usually with varying movement patterns and velocities to maximize uncertainty when sequencing pitch types. However, most pitchers rely on some pitches more than others (particularly depending on the handedness of the batter). Information about a pitcher's pitch usage (and related pitch behavior) is readily available to teams and players, who presumably can then use this information to generate some form of a prior for each pitcher. However, there are cases in which external information can be relayed to the batters about oncoming pitches (e.g., pitch tipping, decoding of signs, and cheating), thus influencing the prior. In this study, we consider a specific case in which the decoding scheme is known. In the fifth game of the 2019 American League Division Series (ALDS)[12], Tyler Glasnow, an elite pitcher for the Tampa Bay Rays, was purportedly tipping his pitches based on how he positioned his glove along the length of his torso. In 2019, Tyler Glasnow threw three pitch types: the 4-seam fastball (67%, 96.9 $\pm$ 0.4 mph), a curve ball (29.3%, 83.5 $\pm$ 0.2 mph), and a change-up (3.5%, 92.9 $\pm$ 0.9 mph); thus, there is a relatively strong prior indicating that two pitches are more probable. However, due to the significantly varying behavior of the fastball and curve ball, batters struggled to hit well against his pitches. In this particular game, it became quickly apparent that batters were able to predict which pitch was being thrown simply based on the position of the glove along his torso (Figure]3B). When comparing the data from this particular game to all other games throughout the 2019 season, batters relied significantly more on the prior when making predictions (Figure 3). Thus, external information (pitch tipping) influences the weight that batters place on the prior and likelihood when making inferences.

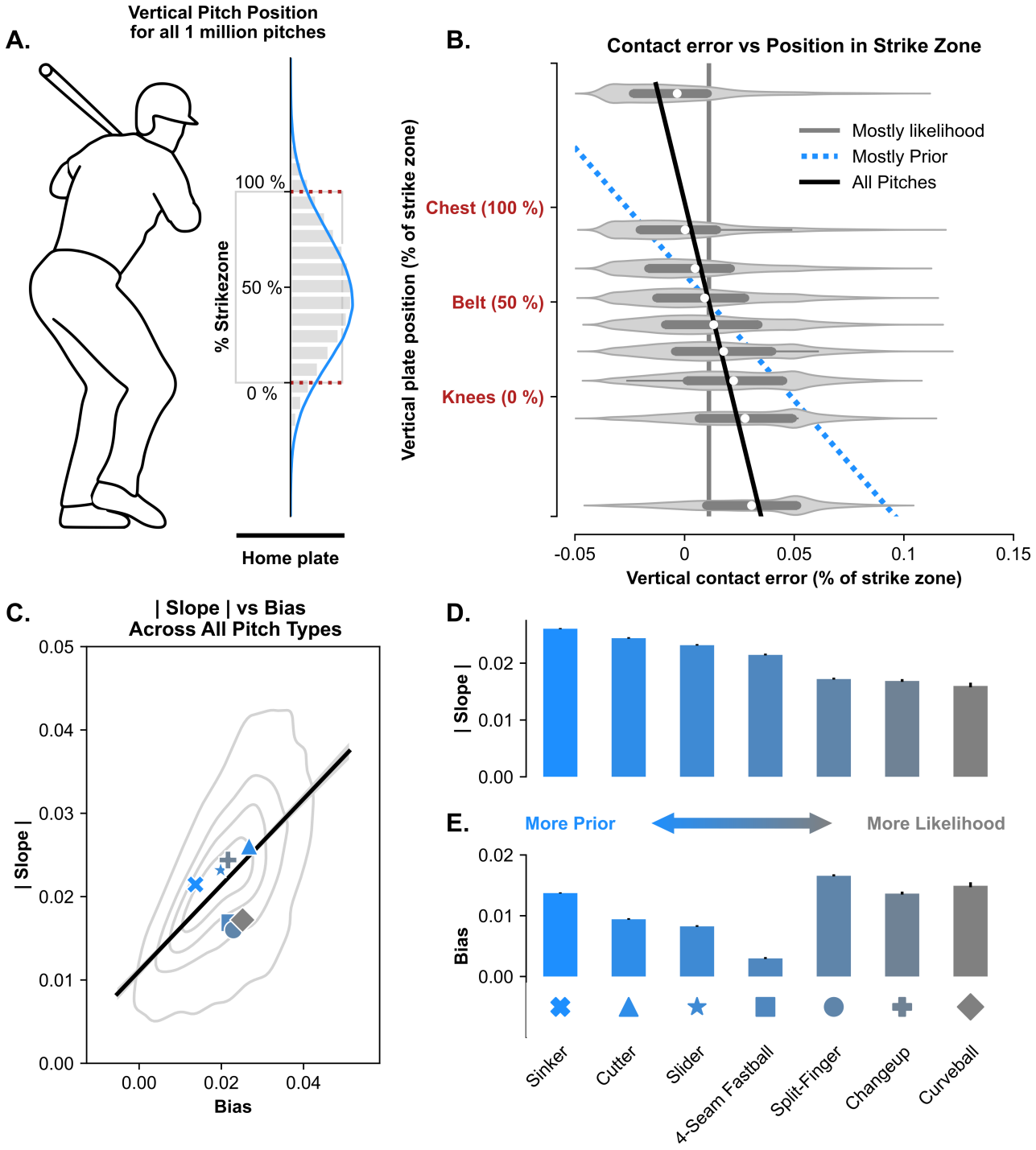

Figure 2. Batters combine prior information with noisy observations to manage pitch uncertainty. **A.** Distribution of vertical pitch location within the strike zone for all 1 million pitches. **B.** Vertical contact error as a function of vertical plate position. The solid gray line indicates a mostly likelihood solution and the dotted line indicates a mostly prior solution. The blue line indicates a Bayes-like solution across all pitch types. **C.** The slope versus bias of each individual pitch type. The distribution around the line indicates the distribution across all individual pitchers. **D.** The slope of each pitch type. **E.** The bias for each pitch type. **Note:** The reported slope is for the regression model when the vertical plate position is treated as the explanatory variable and the contact error is the response. The axes are switched for visual clarity.

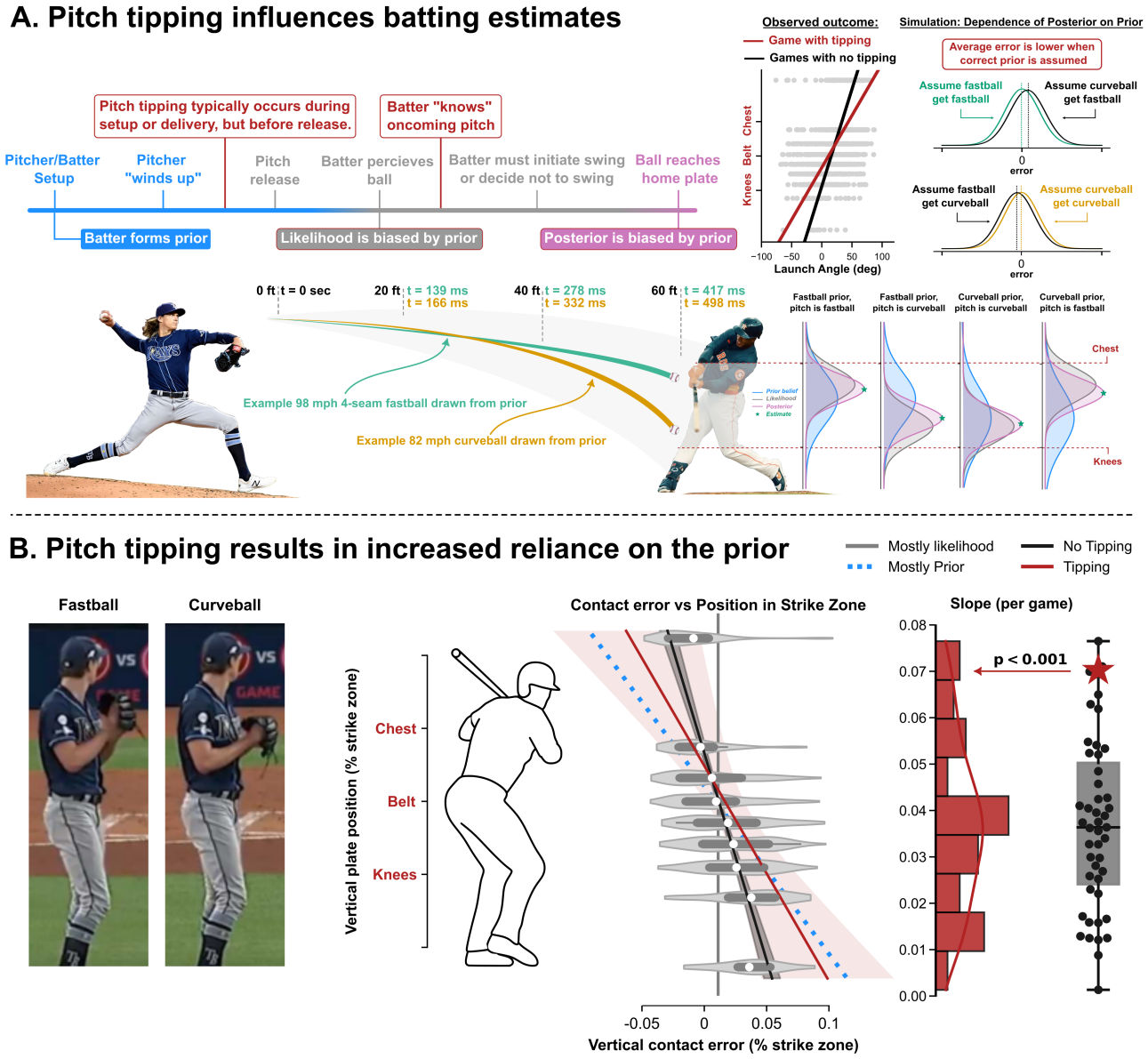

Figure 3. Pitch tipping results in a batting estimate biased toward the prior. A. In the timeline of a single plate appearance, pitch tipping occurs before ball release, either during the setup or during the wind-up. This results in pitchers having prior knowledge about the oncoming pitch. The error between the posterior estimate and the actual ball position is lower when the correct prior is chosen (simulation). **B.** *Left:* In some cases, the posture of a pitcher may give away the planned pitch. Here, when the glove is high, the pitcher is throwing a fastball and when it is lower, the oncoming pitch is a curveball. In this case, we should find more reliance on the now improved prior. *Right:* As predicted, we find that the slope is higher and thus that the batters rely more on the prior.

Pitches with weak priors but high movement uncertainty facilitate dependence on the likelihood

Knuckleball pitch behavior is characterized by erratic and unpredictable movements that make it challenging to predict the ball trajectory due to high movement uncertainty. In the years 2015-2021, there were a total of 3,054 knuckleballs thrown across 8 pitchers. Among these pitchers two pitchers account for 97.6% of all knuckleballs thrown, with one pitcher throwing knuckleballs 83.3% of the time (R.A. Dickey[13]; 2492 total pitches) and the other throwing them 74.7% (Steven Wright[14]; 1210 total pitches). Thus, when either of these pitchers was throwing the ball, there was a strong prior for what pitch the batter was likely to see. However, due to the highly erratic behavior of the knuckleball, the high movement uncertainty renders the prior very weak. As shown in Figure 4D, the knuckleball results in a slope that trends towards zero, indicating a high reliance on the likelihood, despite the potentially strong prior of knowing that the particular pitch type will be thrown. In laboratory settings, this pitch is analogous to a condition in which both the behavior prior and likelihood are drawn from a wide distribution, or a distribution with high uncertainty. In this case, the "weak" prior knowledge (due to high movement uncertainty) results in a posterior estimate having greater reliance on the more "reliable" observed information, despite also being drawn from a wide distribution. This can either be interpreted as the batter ignoring the prior altogether, or simply assuming a flat prior with uniform probability for any location within the strike zone.

Pitches with weak priors but low movement uncertainty facilitate dependence on the likelihood

Another unusual, yet interesting baseball pitch is the Eephus, which is a slow, often looping, curve ball that is thrown at a much lower velocity than all other pitches. The Eephus is a rare pitch, thus resulting in a weak prior for batters when predicting a pitch type (although some pitchers are known for throwing them at a higher rate, it is still uncommon). How then do batters make inferences on a pitch that is so rarely seen? Two distinct traits about the Eephus make it an "easy" pitch to hit: the flight path of an Eephus curveball tends to follow an arc trajectory and the velocity of the Eephus ranges from the low 50 mph to high 60 mph range–velocities otherwise commonly seen by youth players around 12-14 years of age. In sensorimotor experiments, this pitch is analogous to the case of a wide prior and a narrow likelihood and results in subjects relying on the feedback, or likelihood. As expected, our results reveal a slope close to zero (Figure 4), indicating almost complete reliance on the likelihood due to the weak prior and low observation uncertainty. In other words, the slow velocity and arcing movement allow for significantly more time to observe the pitch and thus a near complete reliance on the observation.

Discussion

Here we asked if professional baseball batters use a Bayesian approach when hitting a pitched baseball. We used over 1 million pitches from Major League Baseball games to estimate the error on batted balls and their dependence on the pitch. We found a systematic dependence of contact errors on pitch position within the zone, with each pitch type being accompanied by its own level of uncertainty. Moreover, we found that this is strongly interacting with pitch tipping, where pitcher behavior gives batters better priors. We also observed that when facing highly uncertain or uncommon pitches, the priors are largely ignored resulting in a posterior estimate approximately at the maximum likelihood estimate. These effects are well predicted by a conceptual Bayesian prior-likelihood integration model. Our results suggest that batters rely on prior knowledge and real-time observations (likelihood) when making predictions for batting. In laboratory studies, the observed error, or deviation, can be related back to the true shift imposed by experimenters. When the batter sees the pitcher prior to the throw, they can form an expectation of where the ball will land, e.g. by looking at the pitcher, thus drawing from a distribution known as the prior in Bayesian terms. However, this prior will be different from trial to trial and it cannot be measured directly. The batter's estimate of where the ball will land can then be seen as a way of improving the estimate based on seeing the ball's trajectory. However, we can not measure this prior distribution since it largely depends on what is going on in the batter's brain. But, any ball's position can be seen as traveling towards the mean of the prior and then having an error relative to that. This error is what Bayesian statistics operates upon. As such, we cannot predict the exact dependence of batting errors on the pitched ball trajectories. However, what we can predict is the trend coming from Bayesian statistics: a ball landing lower should lead to an error further up on the ball. This dependence should be represented by a steeper slope if the prior information is better and shallower if the likelihood information is better. For our analysis we have to approximate the calculation of batting contact error, which we consider to be the deviation from the "true shift". Our analysis uses a simple 1D geometrical model that estimates the deviation in the vertical axis based on the launch angle off of the bat. This model does not account for the tilt or fan of the bat, which will have an impact on the contact point on the ball. Additionally, we assume that the attack angle of the bat and the centerline angle of the ball are aligned such that the contact is purely normal to the bat surface. Due to the limitations of our data, we are not able to explicitly address these assumptions in our model; however, we believe that this model is a fair first-order approximation of the bat-ball contact and that the behavior observed in our data is persistent across the millions of pitches incorporated in our analysis. There is no reason to believe that the relevant angles would systematically depend on factors expected to influence the prior. As tracking technology continues to advance, future research can use a full calculation based on precise tracking of the ball and bat. One important question to consider is whether biomechanics alone can explain the results of our analysis. Could it be that our finding that error is related to height within the zone is simply explained by the swing starting at a point where the batter does not have enough time to raise or lower the bat sufficiently? That interpretation seems unlikely to us based on both existing knowledge and our data. It is well established that batters are still able to achieve hits, including home runs, when facing extremely high velocity pitches, even when the ball is high or low. At these fast pitch velocities, the necessary bat velocity required to hit the ball is far too high if the batter waits until the pitch has been fully perceived. In order to hit the ball, the batter must initiate the swing prior to recognizing the ball by relying on some prior belief about the pitch. Moreover, the assumption of the effect being entirely biomechanical can neither explain why tipping leads to what appears to be stronger priors. The effect of a more localized prior is similar to the effect of a batter being weaker, in that both effects predict that larger deviations of ball positions are associated with larger errors. However, the explanation in terms of priors naturally predicts that tipping produces a stronger prior and thus larger deviations and that knuckleballs produce a weaker prior and thus smaller deviations. The biomechanical explanation seems unlikely: Why would batters be weaker in the case of tipping and stronger for

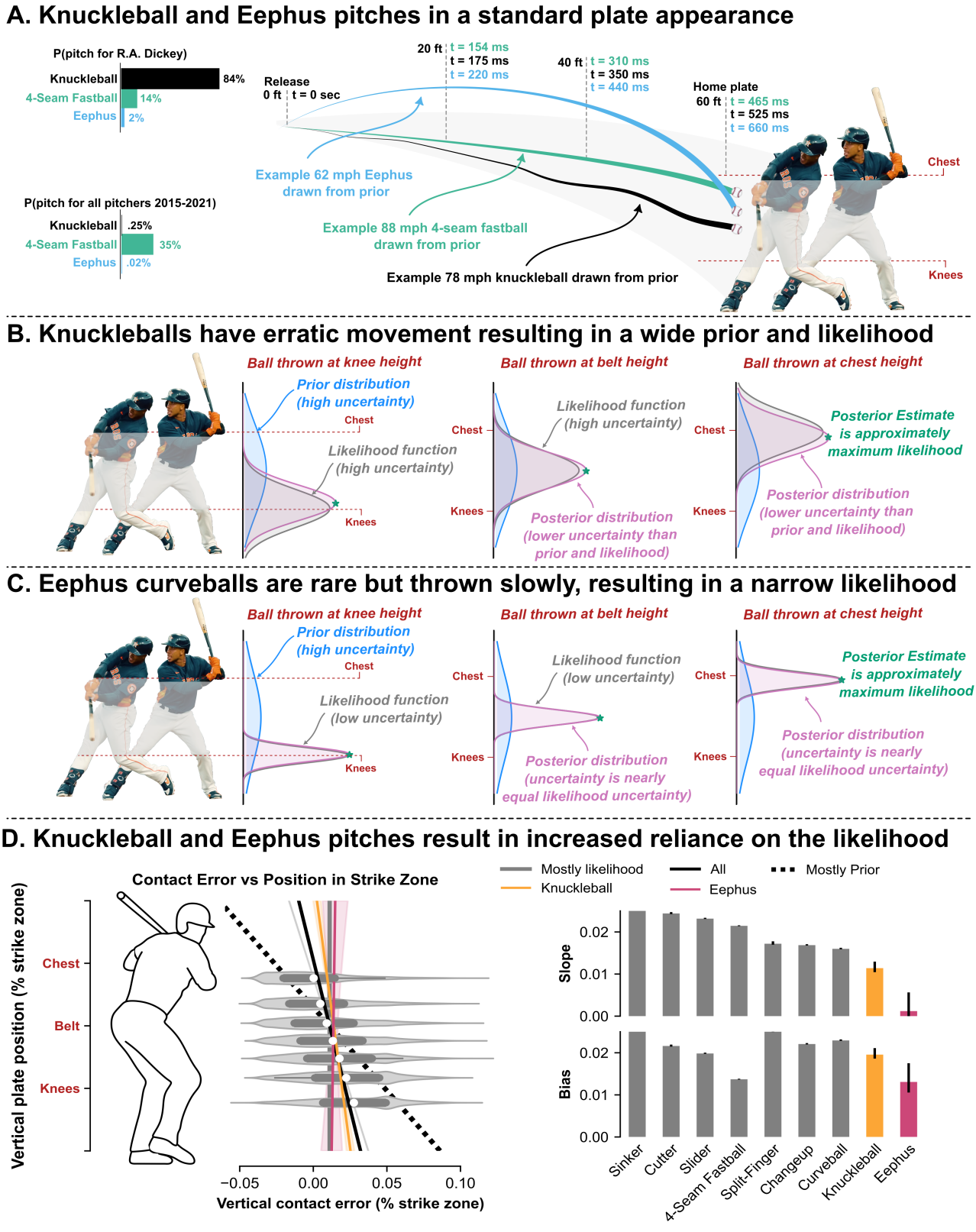

Figure 4. Pitches with weak priors facilitate dependence on the likelihood. A. The movement path of the knuckleball (shown in black) is highly erratic leading to high uncertainty. The Eephus has a slow arcing movement profile (shown in light blue) and occurs very rarely in professional games. **B.** Knuckleballs have a noisy likelihood but a noisier prior, thus it cannot be predicted easily. The posterior estimate is theorized to be approximately the maximum likelihood solution. **C.** The Eephus is a slow pitch and thus easily observable, however, the prior is weak since it is an uncommon pitch. **D.** *Left:* Vertical contact error as a function of vertical plate position. The solid gray line indicates a mostly likelihood solution, the black dotted line indicates a mostly prior solution, and the black line is for all pitches. The orange and pink represent the knuckleball and eephus, respectively. *Right:* The slope and bias of each pitch type. **Note:** The reported slope is for the regression model when the vertical plate position is treated as the explanatory variable and the contact error is the response. The axes are switched for visual clarity.

knuckleballs? As such, the most parsimonious explanation of our data remains the Bayesian one. We do not know precisely what batters are actually trying to do. For example, batters behave differently based on external factors, such as the pitch count, the score, the number of outs, or based on directions provided by the coaches (e.g., "do not swing at the first pitch"). Nonetheless, we believe that these external considerations do not strongly impact the prior or the batter's behavior of integrating the prior and likelihood in a Bayesian manner. Instead, this is likely to only impact the batters interpretation of "optimal" for that specific scenario. For example, a batter attempting to simply get on base (e.g., hit a single) or hit a sacrifice fly to score a runner both result in optimal launch angles not equal to the optimal launch angle for hitting a homerun. While situational batting is important, we do not see how it could explain our data. Explaining baseball behavior in terms of probabilistic ideas has received some attention in the sports science literature. The idea that ball trajectory estimation requires priors and likelihoods has been discussed as parts of the process of batting[15–18] and sports more generally[19]. Our results empirically demonstrate that batters rely on both prior information and observable information when hitting and that the reliability of these information sources can be weighed by their relative uncertainty. Laboratory studies have also shown that the various sources of information that we discuss empirically matter to athletes[15–17,20–25]. The contribution of our study was testing this set of ideas directly on a large database of professional batting in major league baseball games. In the space of sensorimotor integration and motor control, it has been well observed that subjects use priors and noisy feedback based on their respective uncertainties in a way that is consistent with Bayesian statistics. This type of experimental approach, which is typically conducted by introducing some type of visuomotor or dynamic perturbation to a simple movement task, has been iterated in a number of environments, such as in force estimation, cue combination, motor adaptation, and a coin splash game in children. We were inspired by these studies to study baseball to see if this behavior can be observed in the real world. The location where the pitch lands in the strike zone in a way fulfills a similar role to the perturbation in the lab based studies. Just like in lab based studies, there is a prior distribution, with the slight difference that it is not entirely observable for us. In addition, there are factors that influence priors, for example, the case of knuckleballs (relatively flat prior) and tipped pitches (relatively narrow prior) in our study. Moreover, the ball's angle implicitly reveals the estimated position, in the same way as movement patterns reveal implied estimates in lab experiments. With this design we were able to find evidence of Bayesian behavior in professional sports. Our study brings Bayesian movement science into the real world. It opens the path towards computational motor control with big data from other professional sports, such as tennis[26], penalty-kicks in soccer[27], or squash[28]. The emergence of the technology of video based pose-tracking could help to strengthen and enable such analyses. More generally, these findings allow us to observe movements with high quality in the real world. This promises to allow us to bring lab based ideas with their conceptual sophistication to the real world and, ultimately, impart movement science with high ecological validity. The field of motor control needs to bring model based thinking to the real world, and maybe, Bayesian movement control, a large and active field largely confined to laboratories today, will improve athletic performance.

Methods

Data Acquisition

Our analysis was conducted using Python 3.10. All code used for data acquisition and analysis have been made openly available. We used the *pybaseball* library (https://github.com/jldbc/pybaseball) to obtain publicly accessible data from MLB's data clearinghouse[11]. The on-field data are obtained through an acquisition tool known as Statcast[10], which employs high-accuracy, high-speed cameras and radar to track the ball within the field of play, including pitch release velocity, pitch spin rate, ball position within the strike zone, and launch angle (among 91 features returned for each pitch[29]). Our initial query returned a total 4,646,143 pitches for the period of April 1, 2015 through November 1, 2021. We proceeded to exclude games from Spring training or exhibition games since these games are often used for players and teams to experiment with new strategies (e.g., a pitcher exploring a new pitch type, players at new positions, altered mechanics of the swing) and for teams to re-acclimate to playing after the off-season. We then removed pitches in which no contact was made between the bat and ball, including no swing or where the batter would swing and miss.

Estimation of Bayesian Behavior

For our analysis, we used the final vertical position within the strike zone as the representation of the "true shift", or the perturbation drawn from the prior distribution, which we computed by normalizing the true vertical height of the ball (measured in feet) by the height of the strike zone for the batter (computed per pitch). In a laboratory experiment, we would draw the "true shift" from a known prior distribution and then directly measure the deviation from that value based on the error in the batter's estimate. However, since we do not have data that directly track the bat, we infer the contact point based on the launch angle of the batted ball. There are several factors that impact the launch angle of the batted ball, namely the centerline angle of the ball and the attack angle, fan, and tilt of the bat[30]. Due to our inability to track the bat, we therefore assume that contact is purely normal such that there is no fan or tilt and the

attack angle and the center line angle of the ball are aligned at contact. We assume a simple geometrical model, where the contact error, $\mathbf{e}_{\text{contact}}$, is computed as:

$$\mathbf{e}_{\text{contact}} = -\mathbf{r}_{\text{baseball}} \times \Big(\sin(\theta_{\text{optimal}}) - \sin(\theta_{\text{contact}})\Big),$$

where $\mathbf{r}_{\text{baseball}} = 1.45$-inches, $\theta_{\text{optimal}} = 25°$, and θ_{contact} is the launch angle observed when the bat contacts the ball. We then performed a final cleaning by removing extreme outliers in the data resulting from tracking errors. The data were binned into by discretizing the vertical position within the strike zone into 9 bins. We fit an ordinary least squares (OLS) model to fit the error on the batted ball as a function of the position within the strike zone for the aggregate pitch data (all pitch types) as well as for each unique pitch type (minimum number of pitches = 10,000). According to Bayesian statistics, we assume that a slope closer to zero indicates more reliance on the likelihood and an increasing slope indicates more reliance on the prior (Figure 1D). We performed this same analysis for the cases of knuckleballs and eephus pitches for the available number of pitches within the dataset.

Pitch Tipping

For the case of pitch tipping we specifically identified pitches from the game during which the tipping occurred (November 10, 2019) as well as all other pitches thrown by Tyler Glasnow during the 2018-2020 seasons. We followed the same procedure as before by fitting an OLS model for each of the two conditions and used a t-test to determine the statistical significance of the slope during tipping compared to all other pitch appearances.

Data/Code Availability

All code used for data acquisition, data analysis, and figure generating have been made available at https://github.com/KordingLab/Bayesball

Acknowledgements

The authors would like to thank Dr. Brett Mensh for the significant editorial contributions. The authors would like to thank Professor Alan M. Nathan for the helpful discussions regarding the physics of the bat/ball contact model. Thank you to members of the Körding Lab for the feedback and discussion throughout the project.

Author contributions

JAB and KPK conceived of the project together; JAB analyzed the data and created the figures; JAB and KPK wrote the manuscript.

References

1. *Statcast Pitch Arsenals Leaderboard* baseballsavant.com. https://baseballsavant.mlb.com/leaderboard/pitch-arsenals?year=2021&min=100&type=avg_speed&hand= (2022).
2. Gray, R. A model of motor inhibition for a complex skill: Baseball batting. en. *Journal of Experimental Psychology: Applied* **15,** 91–105. ISSN: 1939-2192, 1076-898X. http://doi.apa.org/getdoi.cfm?doi=10.1037/a0015591 (2022) (2009).
3. Körding, K. Decision Theory: What "Should" the Nervous System Do? en. *Science* **318,** 606–610. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.1142998 (2022) (Oct. 2007).
4. Körding, K. P. & Wolpert, D. M. Bayesian integration in sensorimotor learning. en. *Nature* **427,** 244–247. ISSN: 0028-0836, 1476-4687. http://www.nature.com/articles/nature02169 (2022) (Jan. 2004).
5. Körding, K. P. & Wolpert, D. M. Bayesian decision theory in sensorimotor control. en. *Trends in Cognitive Sciences* **10,** 319–326. ISSN: 13646613. https://linkinghub.elsevier.com/retrieve/pii/S1364661306001276 (2022) (July 2006).
6. Vilares, I. & Kording, K. Bayesian models: the structure of the world, uncertainty, behavior, and the brain: Bayesian models and the world. en. *Annals of the New York Academy of Sciences* **1224,** 22–39. ISSN: 00778923. https://onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.2011.05965.x (2022) (Apr. 2011).
7. Arthur, R. *The New Science Of Hitting* en-US. Apr. 2016. https://fivethirtyeight.com/features/the-new-science-of-hitting/ (2022).
8. Berniker, M., Voss, M. & Kording, K. Learning Priors for Bayesian Computations in the Nervous System. en. *PLoS ONE* **5** (ed Brezina, V.) e12686. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0012686 (2022) (Sept. 2010).
9. Körding, K. P., Ku, S.-p. & Wolpert, D. M. Bayesian Integration in Force Estimation. en. *Journal of Neurophysiology* **92,** 3161–3165. ISSN: 0022-3077, 1522-1598. https://www.physiology.org/doi/10.1152/jn.00275.2004 (2022) (Nov. 2004).
10. *Statcast | Glossary* en. https://www.mlb.com/glossary/statcast (2022).
11. *Statcast Search* en_US. https://baseballsavant.mlb.com/statcast_search (2022).
12. *2019 American League Division Series (ALDS) Game 5, Tampa Bay Rays at Houston Astros, October 10, 2019* en. https://www.baseball-reference.com/boxes/HOU/HOU201910100.shtml (2022).
13. *R.A. Dickey Statcast, Visuals & Advanced Metrics | MLB.com* en_US. https://baseballsavant.

mlb.com/savant-player/r-a-dickey-285079 (2022).

14. *Steven Wright Statcast, Visuals & Advanced Metrics | MLB.com* en_US. https://baseballsavant.mlb.com/savant-player/steven-wright-453214 (2022).
15. Gray, R. & Cañal-Bruland, R. Integrating visual trajectory and probabilistic information in baseball batting. en. *Psychology of Sport and Exercise* **36,** 123–131. ISSN: 14690292. https://linkinghub.elsevier.com/retrieve/pii/S1469029217306775 (2021) (May 2018).
16. Gredin, N. V., Bishop, D. T., Broadbent, D. P., Tucker, A. & Williams, A. M. Experts integrate explicit contextual priors and environmental information to improve anticipation efficiency. *Journal of Experimental Psychology: Applied* **24.** Place: US Publisher: American Psychological Association, 509–520. ISSN: 1939-2192 (2018).
17. Gredin, N. V., Broadbent, D. P., Findon, J. L., Williams, A. M. & Bishop, D. T. The impact of task load on the integration of explicit contextual priors and visual information during anticipation. en. *Psychophysiology* **57.** ISSN: 0048-5772, 1469-8986. https://onlinelibrary.wiley.com/doi/10.1111/psyp.13578 (2021) (June 2020).
18. Gredin, N. V., Bishop, D. T., Williams, A. M. & Broadbent, D. P. The use of contextual priors and kinematic information during anticipation in sport: toward a Bayesian integration framework. en. *International Review of Sport and Exercise Psychology,* 1–25. ISSN: 1750-984X, 1750-9858. https://www.tandfonline.com/doi/full/10.1080/1750984X.2020.1855667 (2021) (Dec. 2020).
19. Magnaguagno, L. & Hossner, E.-J. The impact of self-generated and explicitly acquired contextual knowledge on anticipatory performance. en. *Journal of Sports Sciences* **38,** 2108–2117. ISSN: 0264-0414, 1466-447X. https://www.tandfonline.com/doi/full/10.1080/02640414.2020.1774142 (2021) (Sept. 2020).
20. Cañal-Bruland, R., Filius, M. A. & Oudejans, R. R. D. Sitting on a Fastball. *Journal of Motor Behavior* **47.** Publisher: Routledge _eprint: https://doi.org/10.1080/00222895.2014.976167, 267–270. ISSN: 0022-2895. https://doi.org/10.1080/00222895.2014.976167 (2021) (July 2015).
21. Farrow, D. & Reid, M. The contribution of situational probability information to anticipatory skill. en. *Journal of Science and Medicine in Sport* **15,** 368–373. ISSN: 14402440. https://linkinghub.elsevier.com/retrieve/pii/S1440244011004816 (2021) (July 2012).
22. Gray, R. Review: Approaches to Visual-motor Control in Baseball Batting. en. *Optometry and Vision Science* **98,** 738–749. ISSN: 1538-9235, 1040-5488. https://journals.lww.com/10.1097/OPX.0000000000001719 (2021) (July 2021).
23. Higuchi, T. *et al.* Contribution of Visual Information about Ball Trajectory to Baseball Hitting Accuracy. en. *PLOS ONE* **11.** Publisher: Public Library of Science, e0148498. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0148498 (2021) (Feb. 2016).
24. Laby, D. M., Kirschen, D. G., Govindarajulu, U. & DeLand, P. The Effect of Visual Function on the Batting Performance of Professional Baseball Players. en. *Scientific Reports* **9,** 16847. ISSN: 2045-2322. http://www.nature.com/articles/s41598-019-52546-2 (2021) (Dec. 2019).
25. Nasu, D. *et al.* Behavioral Measures in a Cognitive-Motor Batting Task Explain Real Game Performance of Top Athletes. *Frontiers in Sports and Active Living* **2,** 55. ISSN: 2624-9367. https://www.frontiersin.org/article/10.3389/fspor.2020.00055/full (2021) (May 2020).
26. Loffing, F. & Hagemann, N. On-Court Position Influences Skilled Tennis Players' Anticipation of Shot Outcome. *Journal of Sport and Exercise Psychology* **36,** 14–26. ISSN: 0895-2779, 1543-2904. https://journals.humankinetics.com/view/journals/jsep/36/1/article-p14.xml (2021) (Feb. 2014).
27. Wang, Y., Ji, Q. & Zhou, C. Effect of prior cues on action anticipation in soccer goalkeepers. en. *Psychology of Sport and Exercise* **43,** 137–143. ISSN: 14690292. https://linkinghub.elsevier.com/retrieve/pii/S1469029218304709 (2021) (July 2019).
28. Abernethy, B., Gill, D. P., Parks, S. L. & Packer, S. T. Expertise and the Perception of Kinematic and Situational Probability Information. en. *Perception* **30,** 233–252. ISSN: 0301-0066, 1468-4233. http://journals.sagepub.com/doi/10.1068/p2872 (2022) (Feb. 2001).
29. *Statcast Search CSV Documentation* en_US. https://baseballsavant.mlb.com/csv-docs (2022).
30. Kensrud, J. R., Nathan, A. M. & Smith, L. V. Oblique collisions of baseballs and softballs with a bat. *American Journal of Physics* **85.** Publisher: American Association of Physics Teachers, 503–509. ISSN: 0002-9505. https://aapt.scitation.org/doi/10.1119/1.4982793 (2022) (July 2017).