Letter    Editors' Suggestion

# Fitting an active Brownian particle's mean-squared displacement with improved parameter estimation

Maximilian R. Bailey,[1,*] Alexander R. Sprenger,[2,3] Fabio Grillo,[1] Hartmut Löwen,[2] and Lucio Isa[1,†]

[1]*Laboratory for Soft Materials and Interfaces, Department of Materials, ETH Zürich, Vladimir-Prelog-Weg 5, 8093 Zurich, Switzerland*
[2]*Institut für Theoretische Physik II: Weiche Materie, Heinrich-Heine-Universität Düsseldorf, D-40225 Düsseldorf, Germany*
[3]*Institut für Physik, Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany*

(Received 3 August 2022; accepted 3 October 2022; published 22 November 2022)

The active Brownian particle (ABP) model is widely used to describe the dynamics of active matter systems, such as Janus microswimmers. In particular, the analytical expression for an ABP's mean-squared displacement (MSD) is useful as it provides a means to describe the essential physics of a self-propelled, spherical Brownian particle. However, the truncated or "short-time" form of the MSD equation is typically fitted, which can lead to significant problems in parameter estimation. Furthermore, heteroscedasticity and the often statistically dependent observations of an ABP's MSD lead to a situation where standard ordinary least-squares regression leads to biased estimates and unreliable confidence intervals. Instead, we propose here to revert to always fitting the full expression of an ABP's MSD at short timescales, using bootstrapping to construct confidence intervals of the fitted parameters. Additionally, after comparison between different fitting strategies, we propose to extract the physical parameters of an ABP using its mean logarithmic squared displacement. These steps improve the estimation of an ABP's physical properties and provide more reliable confidence intervals, which are critical in the context of a growing interest in the interactions of microswimmers with confining boundaries and the influence on their motion.

DOI: 10.1103/PhysRevE.106.L052602

Overdamped active Brownian motion is often invoked to describe the physics of experimental realizations of active matter [1,2]. The "active Brownian particle's" (ABP) motion is described using Langevin dynamics in the overdamped (inertia-free) regime and consists of an object simultaneously subjected to thermal fluctuations and directed self-propulsion. In this model, the particle moves with a constant velocity $V_0$ in the direction of its internal orientation axis $\hat{\boldsymbol{u}}$, which fluctuates over time due to rotational Brownian motion [3]. Particles therefore travel ballistically over times shorter than the characteristic timescale for rotational diffusion (persistent motion), displaying diffusive motion (with a larger, effective diffusion coefficient) at longer times, as their direction of motion is randomized [4]. This model provides meaningful statistical quantities such as an analytical description for the mean-squared displacement (MSD) of spherical microswimmers, which often shows good agreement with experimental findings [5]. Most analyses in the experimental literature on microswimmers are in fact based on parameters estimated by fitting the sample MSD to the ABP model, extracting particle velocity $V_0$, translational diffusivity $D_T$, and rotational diffusivity $D_R$. In two spatial dimensions, the ABP model prescribes the following expression for the MSD $\langle \Delta \mathbf{r}^2(\tau)\rangle$ as a function of lag time $\tau$ [1,6]:

$$\langle \Delta \mathbf{r}^2(\tau)\rangle = 4D_T\tau + \frac{2V_0^2}{D_R^2}\left(D_R\tau - 1 + e^{-D_R\tau}\right). \quad (1)$$

The standard approach to parameter estimation from a defined model is to use ordinary least-squares (OLS) regression [7,8] following

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{P} (Y_i - f_{i,\theta})^2, \quad (2)$$

where $\hat{\boldsymbol{\theta}}$ is the vector of estimated parameters, $Y_i$ are individual observations from the data set $P$ (here given by the sample's MSD after a given lag time $\tau$), and $f_{i,\theta}$ corresponds to the values of the fitted model [here given by the theoretical prediction; see Eq. (1)]. $\operatorname{argmin}_\theta$ finds the vector $\boldsymbol{\theta}$, which minimizes the objective function. In practice, there are two main strategies to determine the MSD of a population of particles from their coordinates: one can perform either an ensemble average or a time average over the displacements. Ensemble averaging over many particles preserves the statistical independence of the observations and efficiently averages out spurious noise [9], but collecting sufficient statistics in the dilute limit where Eq. (1) holds is experimentally challenging.

Therefore, one often resorts to the calculation of the MSD via time averaging the displacements of a few ABP trajectories followed over time. Moreover, time averaging is advantageous in that it describes the physics of individual microswimmers, whereas studying the EMSD removes information about the heterogeneities present within the system, such as particles displaying atypical motion or changing dynamics within different spatial domains [10]. The time-averaged MSD (TAMSD) of a single particle at a lag time

[*] maximilian.bailey@mat.ethz.ch
[†] lucio.isa@mat.ethz.ch

©2022 American Physical Society

$n\Delta\tau$ is calculated as

$$\text{TAMSD}\,|_{n\Delta\tau} = \sum_{m=1}^{M-n} \frac{\{\mathbf{r}[(n+m)\Delta\tau] - \mathbf{r}(m\Delta\tau)\}^2}{M-n}, \tag{3}$$

where $\mathbf{r}[(n+m)\Delta\tau]$ is the particle position at lag time $n\Delta\tau$ from its previous (reference) position $\mathbf{r}(m\Delta\tau)$, for a trajectory of length $M$. By collecting sufficiently long trajectories, there is the implicit assumption that statistically robust averaging is performed, which is required for accurate parameter estimation with OLS.

However, there are several key assumptions that must be satisfied when using least-squares regression: of these, two can be violated when evaluating the MSD of an ABP. The rotational and time symmetry of a theoretical ABP ensures that consecutive *nonoverlapping* squared displacements are statistically independent, but nonidealities in experimental systems can create hidden correlations, thereby violating the assumption of statistically independent measurements [11,12]. Furthermore, to increase the statistics, one typically evaluates *overlapping* squared displacements when investigating ABPs, which are in fact correlated (see below for further discussion). This impacts the reliability of the estimated confidence intervals, which can become unrealistically narrow. In the case of statistically dependent measurements, confidence intervals for estimated parameters can nevertheless be constructed by fitting the model to bootstrapped datasets from experimental values [11,13].

The second violation is the assumption of homoscedasticity in the error terms of the MSD. There are two sources for heteroscedasticity (nonconstant variance) within the error terms of the MSD with lag time. First, as we show later, the theoretical population variance of an ABP's MSD increases with lag time. Furthermore, the number of data points used to estimate the TAMSD decreases with increasing lag time when evaluating single trajectories, further amplifying the sampling error. These factors, coupled with the presence of localization errors at shorter timescales [14], create a situation where there is an optimal lag time over which the TAMSD of a particle should be evaluated to obtain proper fits of its physical properties [15,16].

To this end, weighted least-squares (WLS) regression is often implemented in order to reduce the dependence of the fit on data points with greater variance, following

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{P} w_{i,\theta}(Y_i - f_{i,\theta})^2, \tag{4}$$

where $\hat{\boldsymbol{\theta}}$ is again the vector of estimated parameters, $Y_i$ are the $P$ data observations, $w_{i,\theta}$ are the weights, and $f_{i,\theta}$ is the model fitted. Here $\operatorname{argmin}_\theta$ now finds the vector $\boldsymbol{\theta}$, which minimizes the weighted objective function. The objective function can be weighted by the inverse of the analytical expression of the population variance (here the variance of the squared displacements) as an estimation of the sample error of the mean [15,17]. The variance of the mean of a random variable $X$, i.e., $\text{E}[X] = \sum_{i=1}^{N} X_i/N$, can be obtained using the variance sum law for uncorrelated variables as

$$\text{Var}[\text{E}[X]] = \text{Var}\left[\sum_{i=1}^{N}\frac{X_i}{N}\right] = \frac{1}{N^2}\sum_{i=1}^{N}\text{Var}[X_i] = \frac{\sigma^2}{N}, \tag{5}$$

where $N$ is the sample size, and $\sigma^2$ is the variance of the random variable $X$. Thus, from Eq. (5), we obtain the following expression for the weights $w_{i,\theta}$:

$$w_{i,\theta} = \frac{1}{\text{Var}[\text{E}[X]]} = \frac{N_i}{\sigma^2_{i,\theta}}, \tag{6}$$

where $N_i$ is the number of statistically independent data points contributing to each observation $i$, and $\sigma^2_{i,\theta}$ is the population variance of each observation $i$, in terms of the fitted values $\boldsymbol{\theta}$.

Nonetheless, the standard approach in the literature is parameter estimation from TAMSDs using unweighted least-squares regression [18]. Additionally, perhaps the most widespread expression that is fitted is the so-called "short-time" MSD of ABPs [19] (7). First proposed by Howse *et al.* for the analysis of Janus catalytic microswimmers [1], the short-time MSD equation approximates the full MSD [Eq. (1)] at an arbitrarily short time lag, typically defined as 10% of the characteristic persistence or rotational diffusion time $\tau_R = 1/D_R$, using a Maclaurin series expansion assuming $\tau/\tau_R \to 0$ [6]

$$\langle \Delta\mathbf{r}^2(\tau)\rangle \sim 4D_T\tau + V_0^2\tau^2. \tag{7}$$

This simplification provides reasonable fits to the experimental TAMSD of single particles under certain conditions, particularly in relation to the extraction of microswimmer velocities [1,18,20–23]. However, care should be taken when fitting this truncated form of the MSD to short experimental trajectories, as it can lead to the spurious detection of velocity in the presence of experimental artifacts [24]. The problems associated with the standard fitting of the truncated form of the MSD were comprehensively demonstrated by Mestre *et al.* [8]. Interestingly, their proposed solution was to expand the Maclaurin series to higher polynomial orders. Nonetheless, we are interested in evaluating the fitting of the full ABP's MSD to the "short-time" regime, as the approximation is simply that: an approximation of a theoretical model.

In this work, we propose multiple approaches to improve the fitting of the full ABP MSD model. We verify the robustness of our approach by comparing it against the "standard" approach of performing unweighted OLS regression on the truncated form of an ABP's MSD at short lag times. We begin by considering the case where $D_T$ and $D_R$ are coupled by the Einstein relation $D_T = d_p^2 D_R/3$ to avoid the introduction of additional fitting parameters and thus allow a fair comparison between the standard approach and our proposed alternatives. In the final section of this study, we then treat $D_R$ as an additional free fitting parameter, corresponding to experimental situations where $D_T$ and $D_R$ are often decoupled. We evaluate the different fitting procedures against simulated ABP trajectories using input values representative of experiments. Specifically, in the coupled case, our ABPs are simulated via Langevin dynamics [25], with an active velocity of $V_0 = 5\ \mu\text{m s}^{-1}$ and diffusivities $D_T = 0.2\ \mu\text{m}^2\ \text{s}^{-1}$ and $D_R = 0.15\ \text{rad}^2\ \text{s}^{-1}$. The simulations are numerically solved at 1 ms

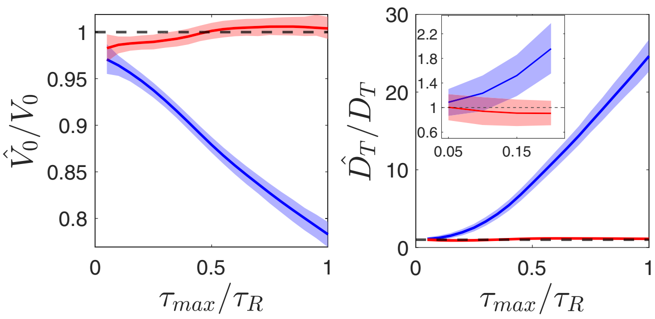

FIG. 1. Parameter estimation from fitting the truncated (blue) and full MSD (red) expression to simulated data (estimates $\hat{V}_0$, $\hat{D}_T$ respectively normalized to the simulation inputs $V_0$, $D_T$). The same trajectory is fitted to increasing maximal lag times $\tau_{max}$, up to the persistence time of an ABP ($\tau_R$). We obtain 95% confidence intervals by bootstrapping. Inset (right): Fits of $D_T$ for short $\tau_{max}$, indicating the rapid deviation from the input simulation value when using the truncated expression.

increments and sampled at 20 frames per second (fps) for 60 s to replicate experimental videos.

Properly applied, there are several advantages to the standard approach of fitting the TAMSD at short timescales. Generally, the scatter of the sample MSD will increase with lag time. This increase not only is caused by the decrease in data points for a trajectory of a given length, but also is due to the growing correlation between sequential observations [see Eq. (3)]. Therefore, the fitting of the MSD to unnecessarily long lag times is generally discouraged [17]. By evaluating the TAMSD over a time period during which the variance does not grow significantly, the effects of heteroscedasticity on parameter estimation are reduced [15]. Nonetheless, the term "short lag times," where the simplified expression holds, is flawed since it is often arbitrarily defined and used in the literature. Furthermore, by eliminating the opportunity to fit $D_R$, the truncated form of Eq. (7) removes characteristic information on the physics of ABPs. Finally, for smaller particles, the characteristic persistence time may be so short that only a few data points can be used to fit the expression, unless experiments are performed at very high frame rates, introducing measurement error and reducing the accuracy of parameter estimation [14].

There are, in fact, further model-specific problems associated with fitting the truncated form of the MSD. As seen in Eq. (2), OLS regression is weighted towards larger values, i.e., MSD values at longer lag times. If left untreated, the fitting of the MSD will therefore be weighted towards the "long-time diffusive" regime of the ABP [4]. Moreover, due to the monotonically growing variance in the error terms of the MSD (discussed below in more detail), this procedure assigns greater importance to more uncertain values, leading to poorer estimates. The effects of these considerations are illustrated by comparing the estimates for $D_T$ and $V_0$ obtained by fitting the truncated and full form of the MSD equation to simulated trajectories (see Fig. 1).

The problems of using Eq. (7) become quickly apparent as the lag times evaluated increase beyond small fractions of the characteristic relaxation time $\tau_R$. As the estimated velocity decreases, the fitted $D_T$ value rapidly increases to over an order of magnitude greater than the simulation input (see Fig. 1, blue). The inverse relationship between $V_0$ and $D_T$ can be understood by their respective contributions to the overall MSD of an ABP. The increasingly diffusive nature of an ABP's motion with time [4] results in an overestimated $D_T$ at the expense of a reduction in the fitted $V_0$. This problem is caused by the absence of the $D_R$-related terms in Eq. (7), which would otherwise result in the crossover to a long-time diffusive regime (see Eq. (1)). In short, due to the systematic errors associated with using Eq. (7) we strongly advise against its use when fitting the MSD of ABPs. To compare the accuracy of our different fitting methods, we use the median symmetric accuracy metric as described in [26]. By evaluating the point estimates over the range of lag times studied, we obtain errors of 14.5% for $\hat{V}_0$ and 799.2% for $\hat{D}_T$ respectively when using the truncated expression for an ABP's MSD.

In contrast, the bootstrapped confidence intervals of the estimated parameters using Eq. (1) more often include the true simulation input values for different maximal lag times $\tau_{max}$ and also converge to reasonable values as the lag time evaluated approaches the characteristic rotational relaxation time $\tau_R$ (see Fig. 1, red). Fitting Eq. (1) also carries the advantage of not assuming a limited short-time regime, enabling the fitting to longer lag times and thus providing more data points for better parameter estimation. Errors on the model parameters estimated are improved to 0.6% and 12.1% for $\hat{V}_0$ and $\hat{D}_T$, respectively. We again emphasize that we do not fit $D_R$ as a free parameter here but instead assume that the Einstein relation $D_T = d_p^2 D_R/3$ holds and fit Eq. (1) accordingly. However, decoupling $D_T$ and $D_R$ better approximates experimental situations where the presence of confining boundaries [27], activity [28–30], or external fields [31] can have a different effect on rotation and translation respectively.

Despite the significant improvement in estimating the physical parameters of an ABP by using the full form of its MSD equation, this operation still does not address underlying statistical issues such as heteroscedasticity of the data. The presence of heteroscedasticity can be clearly observed in the residuals of the fitted ABP model (see Fig. 2, top row, red). One of the most frequently used heuristic approach to address heteroscedasticity is to log transform the data and fit the model's log-transformed analog. Log transforms work particularly well for right skew, constantly positive, and increasing data, such as the case for the ABP's MSD. Studying the "mean logarithmic squared displacement" (MLSD) has previously been suggested to improve the estimation of the distribution of anomalous diffusion coefficients in a population of heterogeneous particles [10].

By fitting the log-transformed (cyan) data, we observe a clear reduction of the heteroscedasticity of the residuals. This provides improved estimated fits and confidence intervals obtained from bootstrapping, and we obtain percentage errors of the point estimates of 0.5% and 2.3% for $\hat{V}_0$ and $\hat{D}_T$ respectively. In Fig. 2 (bottom row), we highlight the improvement in fitting after this simple preprocessing step, evaluating the same trajectory as in Fig. 1 but now with the log-transformed, full MSD ABP fit included as a comparison to the full fit without log transformation. We see both a reduction in the width of the confidence intervals and a smaller difference

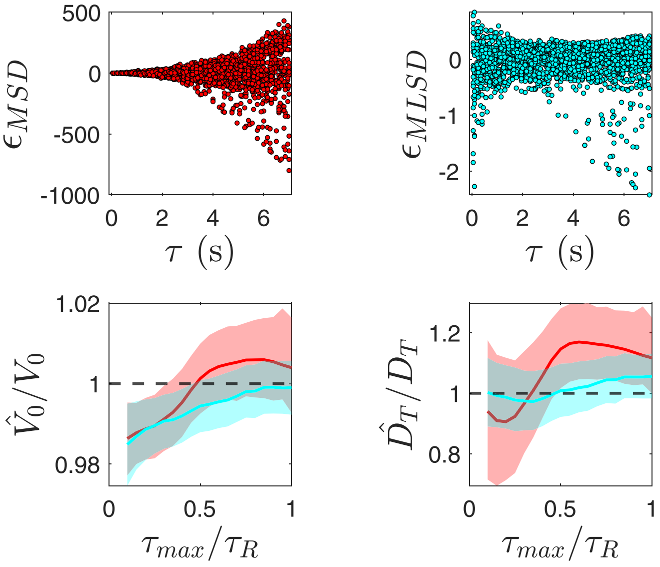

FIG. 2. Top row: Plots of the residuals from the mean squared displacements based on the point estimates in the bottom row for $\tau_{max}/\tau_R = 1$ as a function of lag time $\tau$. Left (red): Residuals of Eq. (1) fitted to unprocessed data (MSD). Right (cyan): Residuals of log[Eq. (1)] fitted to log-transformed data (MLSD). The extent of heteroscedasticity is clearly reduced, as the variance remains relatively constant with $\tau$ after log transformation. Bottom row: Parameter estimation without (red) and with (cyan) log transformation of the data and the model.

between the point estimate and the input simulation values. In particular, the estimates for $D_T$ are notably improved.

As a next step, we turn to WLS regression as a tool for determining the parameters of an ABP. As previously alluded to, within the WLS regression approach, one typically relates the weights to the variance of the expectation value [see Eq. (6)]. Under the assumption that all observations are statistically independent, the variance of the expectation value can be obtained from the population variance itself, using the variance sum law as shown in Eq. (5). Where applicable, we will follow this approach and specify the weights in terms of the theoretical result for the variance of the mean-squared displacement [5,32,33]

$$\begin{aligned}\sigma^2(\tau) &= \langle \Delta\mathbf{r}^4(\tau)\rangle - \langle \Delta\mathbf{r}^2(\tau)\rangle^2 \\ &= 16D_T^2\tau^2 + 16D_T\tau\frac{V_0^2}{D_R^2}\left(D_R\tau - 1 + e^{-D_R\tau}\right) \\ &\quad + \frac{V_0^4}{D_R^4}\bigg(4D_R^2\tau^2 - 22D_R\tau + \frac{79}{2} - \frac{64}{3}D_R\tau e^{-D_R\tau} \\ &\quad - \frac{320}{9}e^{-D_R\tau} - 4e^{-2D_R\tau} + \frac{1}{18}e^{-4D_R\tau}\bigg). \qquad (8)\end{aligned}$$

We note that this result is an exact representation of the variance of the mean only if nonoverlapping squared displacements are considered. For overlapping displacements, a proper analysis requires additional covariance contributions in Eq. (5), describing the correlation between subsequent displacements. In that case, we will still employ Eq. (8), however, as an approximation, and without the contributing term of the number of observations. Equipped with this expression, we can now investigate the presence of heteroscedasticity in an ABP's MSD and attempt to minimize its effects on parameter estimation using WLS regression. As discussed before, we stress that in an experimental context, there might be further hidden correlations between square displacements requiring special consideration, whose evaluation lies beyond the aims of this work. As alluded to above, the TAMSD of particles can be evaluated with one of two different approaches: by determining the overlapping or nonoverlapping particle displacements [see Figs. 3(a) and 3(b)]. Evaluating nonoverlapping squared displacements reduces the correlation between subsequent observations of motion in experimental scenarios and removes it entirely within the framework of the ABP model. However, in this case, the decay in the number of displacements is hyperbolic, decreasing much more rapidly than when overlapping displacements are evaluated [see Fig. 3(c)]. Furthermore, using only nonoverlapping displacements leads to a different sampling of points along the trajectory depending on how many prime factors are present in the number of the time step. These factors lead to a situation where using overlapping displacements typically improves fitting performance and is generally preferable [17].

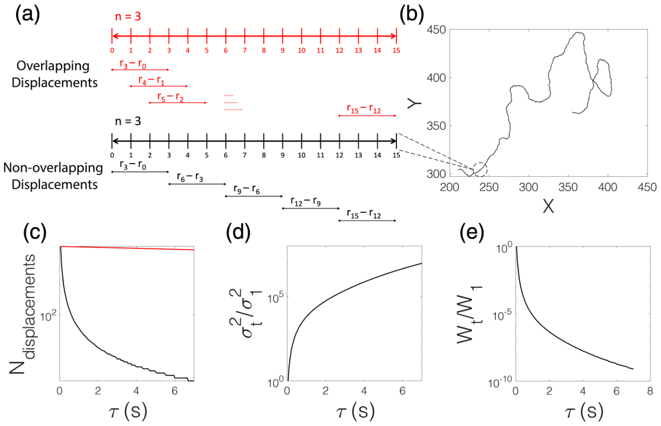

FIG. 3. (a) Definition of overlapping and nonoverlapping displacements from the simulated trajectory shown in (b). (c) Number of displacements as a function of lag time when overlapping (red) and nonoverlapping (black) displacements are evaluated. (d) Normalized variance of the MSD as a function of $\tau$ ($\sigma_1$ is the variance at the shortest lag time $\tau_1 = 0.05$ s), as derived by Eq. (8). (e) Corresponding normalized weight at time $\tau$ ($w_1$ is the weight at $\tau_1 = 0.05$ s) extracted according to Eq. (4) as a function of $\tau$ for the TAMSD of a single particle.

We now discuss the potential benefits of applying the weighting coefficient to minimize the effects of the large and high-variance long lag time values in the objective function [see Eq. (4)]. From Eq. (8), we find that the variance increases with time [see Fig. 3(d)], and combined with the decay in the number of observations [see Fig. 3(c)], we obtain with Eq. (5) a weighting vector that rapidly decays with time [see Fig. 3(e)]. This in turn demonstrates that the low numbers of observations at longer timescales, which inherently have a larger variance due to the nature of the TAMSD, will have a significantly reduced influence on parameter estimation.

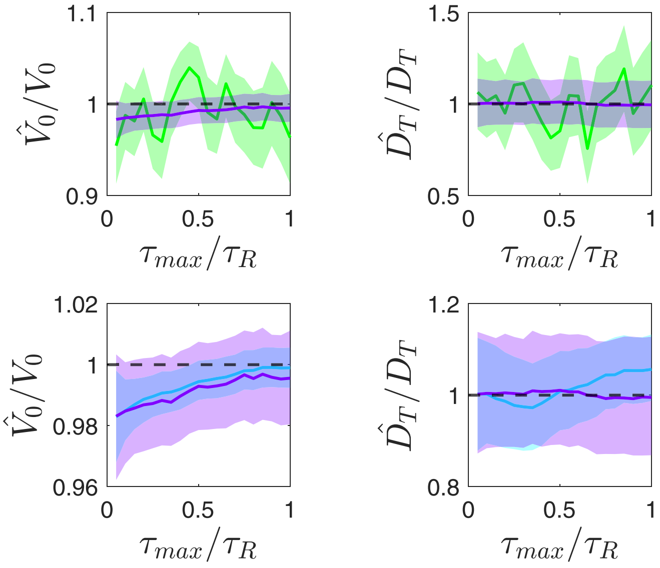

FIG. 4. Top row: Parameter estimation using WLS regression on nonoverlapping (green) and overlapping (purple) displacements. Bottom row: Parameter estimation on overlapping displacements using WLS regression (purple) and the MLSD (cyan).

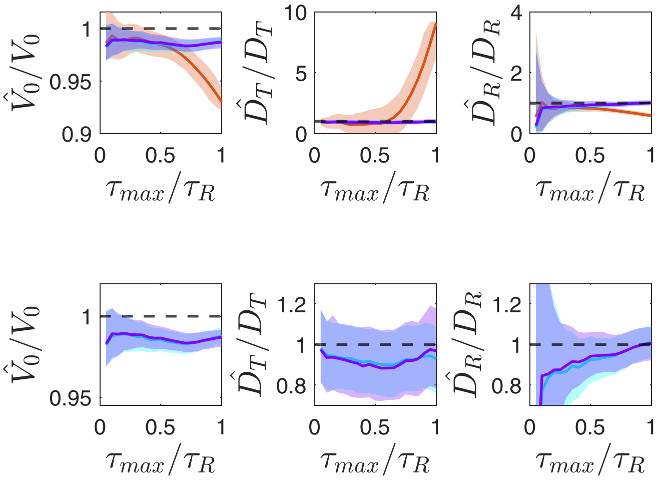

FIG. 5. Parameter estimation of an ABP's MSD where $D_T$ and $D_R$ are uncoupled, using WLS regression (purple), MLSD (cyan), and the third-order truncation of the MSD equation (orange). Top row: Comparison of the three fitting approaches. The truncated expression clearly performs worse, particularly at larger $\tau_{max}$ (see the estimates for $D_T$ and $D_R$). Bottom row: Only the MLSD and WLS regression are represented for better visualization.

We now fit the TAMSD of a single particle using WLS regression, beginning with the analysis of nonoverlapping displacements (see Fig. 4, top row, green). We obtain percentage errors of 2.0% and 5.9% for $\hat{V}_0$ and $\hat{D}_T$, respectively, for the point estimates. We note a significant instability in the point estimates and confidence intervals, particularly for $\hat{D}_T$, in direct comparison to the fits obtained with the MLSD. Therefore, we also evaluate the performance of WLS regression on overlapping displacements, noting that the underlying assumption of statistically independent observations no longer holds (see Fig. 4, top row, purple). We again highlight here that the variance sum law no longer holds, and therefore we weight the objective function for overlapping displacements using only Eq. (8). Comparing the overlapping to the nonoverlapping case, we find that the resulting confidence intervals and point estimates for WLS regression are much narrower and less subject to fluctuations. Under these conditions, we observe percentage errors of 0.7% and 0.5% for $\hat{V}_0$ and $\hat{D}_T$, respectively. We expect this discrepancy arises, in large part, from the statistical issues associated with evaluating nonoverlapping displacements, as described in [17]. Motivated by the improved parameter estimation, we continue to evaluate WLS regression using overlapping displacements for the rest of this work.

We now compare the performance of the MLSD and WLS regression for parameter estimation from overlapping displacements (see Fig. 4, bottom row). Although the resulting confidence intervals are broader for the WLS regression than for the MLSD, we note that in the former case the estimate for $D_T$ is more stable, and the true simulation input parameters are included for all values of $\tau_{max}$. We conclude that for a two-parameter fit, where $D_T = d_p^2 D_R/3$, the estimates obtained from WLS regression and OLS regression of the MLSD are similar.

So far, we have considered only particles satisfying the ideal condition where $D_T$ and $D_R$ are related by the Einstein relationship for freely diffusing spherical particles. However, in many situations, e.g., when in proximity with a solid wall, $D_T$ and $D_R$ are likely to be decoupled [2,27–30], and it is therefore important, in most experimental realizations of ABPs, to fit these parameters separately. We account for these circumstances by modifying the value of $D_T$, while keeping the same value of $D_R$ in our simulations. In particular, we modify the translational diffusivity by applying Faxen's correction factor to $D_T$, as if to mimic the presence of a solid wall 250 nm away from the particle surface [34]. This correction approximately reduces the theoretical $D_T$ value we initially used by half.

In Fig. 5 we compare the performance of the MLSD and WLS regression approaches when estimating the parameters $V_0$, $D_T$, and $D_R$ (blue and purple, respectively). For the MLSD, we determine percentage errors of 1.5%, 7.8%, and 7.9% for $\hat{V}_0$, $\hat{D}_T$, and $\hat{D}_R$, respectively, for the point estimates across all the lag times evaluated, while for the WLS regression we obtain corresponding errors of 1.4%, 8.7%, and 5.9%. We also study the truncated MSD equation expanded to third order, as outlined in [8] (Fig. 5, top row, orange). This expression is obtained by evaluating the Maclaurin series expansion of Eq. (1) to the third order

$$\langle \Delta \mathbf{r}^2(\tau) \rangle \sim 4D_T\tau + V_0^2\tau^2 - \frac{V_0^2}{3\tau_R}\tau^3. \tag{9}$$

We find that as before, the truncated form of the full MSD equation is not able to satisfactorily capture the input simulation parameters, an effect which is particularly noticeable for $D_T$ as $\tau_{max}$ increases, as previously observed in Fig. 1. We determine percentage errors of 1.6%, 32.2%, and 24.4% for $\hat{V}_0$, $\hat{D}_T$, and $\hat{D}_R$ respectively. We note the use of the median function in the median symmetric accuracy metric [26], and the effect this has on the measured accuracy relative to the instability observed in Fig. 5 (top row, orange).

When evaluating overlapping displacements using WLS regression and the MLSD, we note a remarkable overlap in both the point estimates and confidence intervals (see Fig. 5, bottom row). This observation indicates that both the log transformation and weighting of the data have a similar effect on addressing the heteroscedasticity present in an ABP's MSD. In both instances, we also note the instability of the short-time estimates for $D_R$, which is unsurprising given the independence of the MSD from $D_R$ at short lag times [see Eq. (7)].

In conclusion, the ABP model provides a useful framework to study the motion of microswimmers and extract meaningful physical properties from mean quantities. However, "blind" fitting of MSDs can affect results, as hidden correlations may arise in experimental systems. Therefore, we recommend constructing confidence intervals by bootstrapping in almost all experimental situations. We additionally always advise against the use of the truncated form of the MSD equation. Further steps beyond fitting to short lag times should also be taken to treat the heteroscedasticity of an ABP's MSD. In particular, we find that log transforming the data before fitting the MLSD equation outperforms standard approaches used in literature, and provides similar estimates as WLS regression using the theoretical variance of an ABP's MSD. With this approach, overlapping displacements can be evaluated, significantly increasing the amount of data available. Furthermore, the simplicity of fitting log-transformed data to shorter lag times should assist in its widespread uptake. We nevertheless stress that we have studied simulated data of an ideal, noninteracting ABP model, neglecting, e.g., the presence of torque in the Langevin force balance [28,35], a situation that is often observed in experiments due to nonsymmetric surface modification [36] or shape [37], which can significantly affect the fitting of model parameters. Signatures for an angular propulsion velocity should therefore be additionally investigated when analyzing experimental trajectories, and its effect duly included in the fits. We have also not treated the effect of ABP speed on the coupling between $D_T$ and $V_0$ [38] and experimental errors from static and dynamic localization errors [10,15,16,39]. These are nevertheless critical factors which should be considered when designing experiments and analyzing data.

The authors thank Dr. S. Ketzetzi and Dr. B. ten Hagen for various insightful discussions.

Author contributions are defined based on the CRediT (Contributor Roles Taxonomy). Conceptualization: M.R.B., F.G. Formal analysis: M.R.B., A.R.S. Funding acquisition: L.I. Investigation: M.R.B., A.R.S. Methodology: M.R.B., F.G. Software: M.R.B. Supervision: F.G., L.I., H.L. Validation: M.R.B., A.R.S. Visualization: M.R.B. Writing: original draft: M.R.B., A.R.S., L.I. Writing: review and editing: M.R.B., A.R.S., H.L, L.I.

[1] J. R. Howse, R. A. Jones, A. J. Ryan, T. Gough, R. Vafabakhsh, and R. Golestanian, Phys. Rev. Lett. **99**, 048102 (2007).

[2] K. Dietrich, D. Renggli, M. Zanini, G. Volpe, I. Buttinoni, and L. Isa, New J. Phys. **19**, 065008 (2017).

[3] B. ten Hagen, S. Van Teeffelen, and H. Löwen, J. Phys.: Condens. Matter **23**, 194119 (2011).

[4] C. Bechinger, R. Di Leonardo, H. Löwen, C. Reichhardt, G. Volpe, and G. Volpe, Rev. Mod. Phys. **88**, 045006 (2016).

[5] X. Zheng, B. ten Hagen, A. Kaiser, M. Wu, H. Cui, Z. Silber-Li, and H. Löwen, Phys. Rev. E **88**, 032304 (2013).

[6] H. Löwen, J. Chem. Phys. **152**, 040901 (2020).

[7] A. van den Bos, *Parameter Estimation for Scientists and Engineers* (John Wiley & Sons, New York, 2007).

[8] R. Mestre, L. S. Palacios, A. Miguel-López, X. Arqué, I. Pagonabarraga, and S. Sánchez, arXiv:2007.15316 (2020)

[9] F. Novotný and M. Pumera, Sci. Rep. **9**, 13222 (2019).

[10] E. Kepten, I. Bronshtein, and Y. Garini, Phys. Rev. E **87**, 052713 (2013).

[11] K. Fogelmark, M. A. Lomholt, A. Irbäck, and T. Ambjörnsson, Sci. Rep. **8**, 6984 (2018).

[12] H. Qian, M. P. Sheetz, and E. L. Elson, Biophys. J. **60**, 910 (1991).

[13] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall/CRC, Philadelphia, 1994).

[14] E. Kepten, A. Weron, G. Sikora, K. Burnecki, and Y. Garini, PLoS ONE **10**, e0117722 (2015).

[15] X. Michalet, Phys. Rev. E **82**, 041914 (2010).

[16] J. Devlin, D. Husmeier, and J. A. Mackenzie, Phys. Rev. E **100**, 022134 (2019).

[17] M. J. Saxton, Biophys. J. **72**, 1744 (1997).

[18] X. Arqué, A. Romero-Rivera, F. Feixas, T. Patiño, S. Osuna, and S. Sánchez, Nat. Commun. **10**, 2826 (2019).

[19] W. Wang and T. E. Mallouk, ACS Nano **15**, 15446 (2021).

[20] A. M. Pourrahimi, K. Villa, C. L. Manzanares Palenzuela, Y. Ying, Z. Sofer, and M. Pumera, Adv. Funct. Mater. **29**, 1808678 (2019).

[21] V. Sridhar, F. Podjaski, J. Kröger, A. Jiménez-Solano, B. W. Park, B. V. Lotsch, and M. Sitti, Proc. Natl. Acad. Sci. U. S. A. **117**, 24748 (2020).

[22] S. Ketzetzi, J. De Graaf, R. P. Doherty, and D. J. Kraft, Phys. Rev. Lett. **124**, 048002(R) (2020).

[23] M. R. Bailey, N. Reichholf, A. Flechsig, F. Grillo, and L. Isa, Part. Part. Syst. Charact. **39**, 2100200 (2021).

[24] G. Dunderdale, S. Ebbens, P. Fairclough, and J. Howse, Langmuir **28**, 10997 (2012).

[25] A. Callegari and G. Volpe, Flowing Matter **1**, 211 (2019).

[26] S. K. Morley, T. V. Brito, and D. T. Welling, Space Weather **16**, 69 (2018).

[27] A. J. Goldman, R. G. Cox, and H. Brenner, Chem. Eng. Sci. **22**, 637 (1967).

[28] S. Ebbens, R. A. L. Jones, A. J. Ryan, R. Golestanian, and J. R. Howse, Phys. Rev. E **82**, 015304(R) (2010).

[29] S. Das, A. Garg, A. I. Campbell, J. Howse, A. Sen, D. Velegol, R. Golestanian, and S. J. Ebbens, Nat. Commun. **6**, 8999 (2015).

[30] J. Simmchen, J. Katuri, W. E. Uspal, M. N. Popescu, M. Tasinkevych, and S. Sánchez, Nat. Commun. **7**, 10598 (2016).

[31] A. R. Sprenger, M. A. Fernandez-Rodriguez, L. Alvarez, L. Isa, R. Wittkowski, and H. Löwen, Langmuir **36**, 7066 (2020).

[32] C. Kurzthaler and T. Franosch, Soft Matter **13**, 6396 (2017).

[33] F. J. Sevilla and P. Castro-Villarreal, Phys. Rev. E **104**, 064601 (2021).

[34] S. Ketzetzi, J. de Graaf, and D. J. Kraft, Phys. Rev. Lett. **125**, 238001 (2020).

[35] S. van Teeffelen and H. Löwen, Phys. Rev. E **78**, 020101(R) (2008).

[36] X. Wang, M. In, C. Blanc, A. Würger, M. Nobili, and A. Stocco, Langmuir **33**, 13766 (2017).

[37] F. Kümmel, B. ten Hagen, R. Wittkowski, I. Buttinoni, R. Eichhorn, G. Volpe, H. Löwen, and C. Bechinger, Phys. Rev. Lett. **110**, 198302 (2013).

[38] E. M. Tang and P. T. Underhill, Langmuir **34**, 10694 (2018).

[39] T. Savin and P. S. Doyle, Biophys. J. **88**, 623 (2005).