Euro. Jnl of Applied Mathematics (2018), *vol.* 29, *pp.* 1141–1177.
doi:10.1017/S0956792518000220

A point process model for generating biofilms with realistic microstructure and rheology†

JAY ALEXANDER STOTSKY, VANJA DUKIC and DAVID M. BORTZ

Department of Applied Mathematics, University of Colorado, Boulder, CO 80309-0526, USA
emails: Jay.Stotsky@colorado.edu, Vanja.Dukic@colorado.edu, dmbortz@colorado.edu

(*Received 16 July 2017; revised 21 April 2018; accepted 23 April 2018; first published online 16 May 2018*)

Biofilms are communities of bacteria that exhibit a multitude of multiscale biomechanical behaviours. Recent experimental advances have led to characterisations of these behaviours in terms of measurements of the viscoelastic moduli of biofilms grown in bioreactors and the fracture and fragmentation properties of biofilms. These properties are macroscale features of biofilms; however, a previous work by our group has shown that heterogeneous microscale features are critical in predicting biofilm rheology. In this paper, we use tools from statistical physics to develop a generative statistical model of the positions of bacteria in biofilms. Specifically, the model is a type of pairwise interaction model (PIM). We show through simulation that the macroscopic mechanical properties of biofilms depend on the choice of microscale spatial model. A key finding is that uniform and non-uniform sets of points lead to differing mechanical properties. This distinction appears not to have been previously considered in mathematical biofilm literature. We also found that realisations of a biologically informed PIM have realistic *in silico* mechanical properties, and have statistical properties that closely match experimental data. We also note that a Poisson spatial point process of suitable number density also yields realistic mechanical properties, but that the spatial distribution of points does not reflect those occurring in our experimentally observed biofilm.

Key words: Biofilms, Biomechanics, Nonparametric Density Estimation, Spatial Stochastic Processes

1 Introduction

Biofilms are complex, multiorganism communities of bacteria. They are abundant in nature and grow readily in many industrial systems, where they often cause maintenance and safety issues. Measures to mitigate or remove biofilms, though costly, are often necessary in the design and operation of many industrial systems [16,22]. The demand for better biofilm control strategies and the causative role of biofilms in bacterial infections has inspired the development of numerous mathematical biofilm models [17,22,25,31,43, 46,50,54–56]. Despite a 30-year history, it is only recently that the validation of models by comparison with empirical data has become a focus in computational biofilm studies. Recent works have shown agreement between experiment and simulation regarding the

† This work was supported in part by the National Science Foundation grants PHY-0940991 and DMS-1225878 to DMB, and by the Department of Energy through the Computational Science Graduate Fellowship program, DE-FG02-97ER25308, to JAS.

frequency dependent dynamic moduli [43], the ranges of certain spring constants [50], and the rate of disinfection of biofilms in response to antimicrobial substances [56]. However, this recent progress has also led to new questions.

Due to the complex behaviour of biofilms, even state-of-the-art models rely on many assumptions. Material parameter values, biofilm morphology, and properties of a viscoelastic network of fibres that link bacteria together (known as the extracellular polymeric substance) are often specified heuristically. Furthermore, even when parameters are informed by experimental results, there are difficulties. For instance, the experimental measurements of biofilm properties may differ by orders of magnitude between studies depending on the specifics of the conditions under which biofilms were grown and tested [36].

The assumptions noted before involve microstructural properties. Although, some microstructural properties have been measured [13,42], little work has been done to elucidate the influence of microstructural properties on macroscale behaviours. Along these lines, there are two main goals to this paper. The first is to introduce a microstructural description of *Staphylococcus epidermidis* biofilms by estimating certain fundamental statistical characteristics from experimental data. The second is to numerically demonstrate the efficacy of first- and second-order summary statistics for generating data with similar material properties as experimental data sets. The similarity of material properties is tested through simulation using the *heterogeneous rheology Immersed Boundary Method* (hrIBM) [25,43].

1.1 Biofilm point data: Description and pre-processing

Here, we provide a summary of the techniques used to (1) obtain data on the locations of bacteria *in situ* and (2) process such data to make it suitable for statistical analysis. Detailed descriptions of these steps can be found in [42] and [36] and the references therein.

Measuring properties of live biofilms is difficult due to the micrometer scale thickness of biofilms and empirical observations that they deform upon removal from an aqueous environment. In [42], high resolution confocal laser scanning microscopy (CLSM) is used to image biofilms because it is nondestructive, can be applied to live biofilms, and resolves the biofilm up to features of about 0.5 μm in size—small enough to capture individual bacteria. With CLSM, a series of images, each vertically offset by a small amount from the previous, are layered together to provide a three-dimensional (3-D) data set.

The raw CLSM data forms a 3-D grey-scale image of a biofilm. Using the image analysis methods described in [10], the centroid of each individual bacteria can then be obtained to very high accuracy compared to the size of a bacteria. We use the list of bacteria centroids as the data for our model. An image of a biofilm point data set is depicted in Figure 1.

This data set contains a list of 3-D coordinates of the centres of mass of about 4,000 bacteria from live biofilms. Moving from $z = 0$ to $z \approx 10$, takes us from the biofilm-substrate interface to the biofilm-fluid interface. In this paper, results from this data set

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

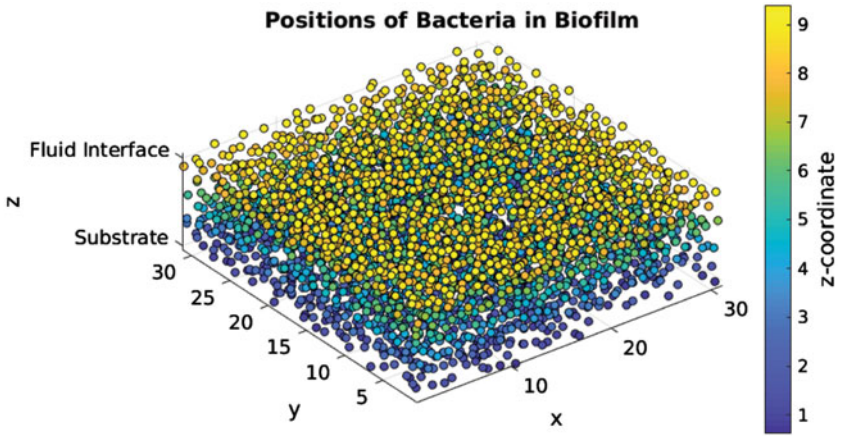

FIGURE 1. An experimental data set from a live biofilm. The data set is approximately 30 μm × 30 μm × 10 μm in size and consists of the centres of mass of 3,981 live bacteria. This data was obtained from high-resolution confocal microscopy images. The techniques used to obtain this data are discussed in [42]. The colouring is for ease of viewing and simply corresponds to the z-coordinate of each bacterium.

and three additional data sets obtained through the same techniques serve as a basis on which our statistical analysis is conducted.

1.2 Overview of mathematical approach

To begin, we develop a point process model of the positions of bacteria in a biofilm that is parametrised by quantities derived from the experimental data. In particular, using estimators of the number density and pair correlation function (PCF), we used a numerical method to solve an integral equation, known as the Ornstein–Zernike equation, and employed an approximate relation known as the hypernetted-chain equation (HCE) to obtain a pairwise potential energy function. This pair potential, along with an external potential, whose derivation from an integral relation is discussed in Section 2.2, is used to compute the value of an unnormalised probability density function associated with the configuration of points. We then generated biofilm realisations by using a Metropolis–Hastings (MH) algorithm with that is constructed to converge to a certain probability distribution discussed in Section 2. Finally, in Section 6, we use the hrIBM to compare material properties of the statistically generated biofilm samples to experimental data.

We also note that the overall procedure bears similarity to another recent article. In [15], a statistical model of red blood cell cytoskeletal structure parametrised by empirical data is introduced, and immersed boundary method simulations are used to estimate biomechanical properties of the model, which are quantitatively compared to experimental data.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

1.3 Organisation of the paper

In Section 2.1, we introduce the statistical model for the positions of bacteria in a biofilm. The model, known as a pairwise interaction model (PIM), is designed to accurately replicate first- and second-order spatial statistics. Similar types of models are frequently studied in the statistical mechanics of fluids, and much of our discussion of the model follows the classic texts [26, Section 5] and [33, Section 6].

In Section 3, we detail the non-parametric estimators of summary statistics of finite point processes that are necessary for the computation of energy functions in the PIM model. These estimators are based on those analysed in [29, 30, 33, 44, 45, 47]. In Section 4, we discuss the numerical solution of the integral equations used to obtain the energy functions from estimable spatial statistics of the data.

From the PIM, an unnormalised probability density associated with the configuration of bacteria in the biofilm is obtained and used in Markov Chain Monte Carlo (MCMC) simulations to generate 'artificial' biofilms [33, 53]. Details can be found in Section 5. Once realisations of the PIM have been obtained, the material properties of the artificial biofilms generated by the PIM, along with some previous biofilm generation models (e.g., [2, 46]), can be compared to those of the experimental data obtained through high resolution CLSM. This comparison is achieved through simulations using the hrIBM to estimate the dynamic moduli of the resulting biofilms. In Section 7, we discuss the results, motivate some future research directions, and suggest potential improvements.

2 A pairwise interaction model of bacterial biofilms

In this section, we first introduce the statistical model and then discuss how parameters in the model are obtained through the use of statistical estimators and the solution of certain integral equations. In constructing a model of the positions of bacteria in a biofilm, we require that the model is physically realistic. In particular, the model should be able to accurately account for volume exclusion effects due to the finite size of bacteria, regularity, and spatial coordination in the positions of bacteria, and possess statistical characteristics that match those computed from the experimental data. Furthermore, since relatively few attempts have been made to analyse such biofilm data, we require a model with parameters that may be estimated empirically. Lastly, to enhance the clarity of this work, we have compiled in Table 1 a list of definitions of all frequently used variables and functions.

A fairly general class of models that can account for these properties are the Markov point process models [33, Section 6]. Among the various types of Markov point process models, PIMs are an attractive choice for this application. They are among the simplest Markov point process models, requiring only the specification of an external potential and a pairwise potential, and there is a well-established theory for obtaining such potentials empirically from the data with minimal assumptions.

The probability density associated with a point configuration of n points, denoted $\boldsymbol{X}$, for a general PIM is of the form [26, Section 2]

$$f^{(n)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{1}{Z_n} \exp\left[-\sum_{i=1}^{n} \phi(\boldsymbol{x}_i) - \sum_{i=1}^{n} \sum_{j=i+1}^{n} v(\boldsymbol{x}_i, \boldsymbol{x}_j) \right], \tag{2.1}$$

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

Table 1. *Definitions of commonly used symbols*

Symbol	Definition
W	Borel set containing experimental data
$\boldsymbol{r}$, (x, y, z)	Generic point in $\mathbb{R}^3$
$\boldsymbol{x}$	Point contained in a point process realisation
$\boldsymbol{r}^n$	Collection of n points in $\mathbb{R}^3$
$\boldsymbol{X}$	Realisation of a point process
$\boldsymbol{\Phi}$	Generic point process
$v(B)$	Lebesgue measure of a set, B
$\rho^{(2)}(\boldsymbol{r}_1, \boldsymbol{r}_2)$	Second-order factorial moment density
$g(\boldsymbol{r}_1, \boldsymbol{r}_2)$	Second-order correlation function
$h(\boldsymbol{r}_1, \boldsymbol{r}_2)$	Indirect correlation function, $h^{(2)} = g^{(2)} - 1$
$f^{(n)}(\boldsymbol{r}^n)$	nth order probability density
$\rho(\boldsymbol{r})$	Number density (also known as the intensity), equivalent to $\rho^{(1)}(\boldsymbol{r})$
$\hat{q}$	Estimator for quantity q
$\tilde{q}$	Hankel transform of quantity q
$\mathbb{E}[q]$	Expectation of q
$k_b(r)$	Smoothing kernel with support of radius b
$c_{b,W}(z)$	Edge correction factor
$c(\boldsymbol{r}_1, \boldsymbol{r}_2)$	Direct correlation function
$\phi(\boldsymbol{r})$	External potential
$v(\boldsymbol{r}_1, \boldsymbol{r}_2)$	Pair potential
$\boldsymbol{e}_x$	Unit vector in x-coordinate direction
$H_k(r)$	Nearest-neighbour distribution function

The nth-order probability density when integrated over a product set, $B_1 \times B_2 \times \ldots B_n$, with $B_i \cap B_j = \emptyset$, gives the probability that for each $\boldsymbol{r}_i$, there exists exactly one B_k such that $\boldsymbol{r}_i \in B_k$ for $i \in [1, n]$.

where ϕ is known as an external potential, v a pairwise potential[1], and the normalisation factor, Z_n the canonical partition function. The probability density, $f_n(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) d\boldsymbol{x}_1, \ldots, d\boldsymbol{x}_n$ can be interpreted as the probability, given n distinct points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of there being a single point in each of the regions of volume $d\boldsymbol{x}_i$ centred at $\boldsymbol{x}_i$.

Variations in the external potential are closely related to variations in the number density of bacteria in various parts of the domain. For instance, given data in a bounded region, outside of the bounded region, we could set $\phi(\boldsymbol{r}) = \infty$ so that $f^{(n)}$ tends to zero if any points are outside the domain. This property will be useful in Section 5.

For our case, the domain, denoted W is the smallest rectangular solid containing all data points. In particular, we set the minimum x-, y-, and z-coordinates to 0, and the maximum coordinates correspond to the maximum of each coordinate over all data points within the data sets. The assumption of W being a rectangular solid seems justified by the fact that the data is relatively homogeneously distributed throughout W (although there is some number density variation, see Section 3).

[1] In the statistical physics community, it is customary to include a temperature dependence on the energy as $\beta v(\boldsymbol{r}_1, \boldsymbol{r}_2)$ for the pair potential (or external potential), where β is an inverse temperature [26, Section 2]. In this paper, we are not concerned with thermal effects, so we redefine $v \leftarrow \beta v$ and $\phi \leftarrow \beta \phi$. Furthermore, without some assumption on the 'temperature' of the biofilm, β and v or ϕ are not independently estimable quantities.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

The form of the pair potential is crucial to accurately modelling interactions between bacteria. Since we do not know of previous studies on the statistical generation of biofilm data, we take an empirical approach to obtaining a pair potential. Other possible choices we could have made include Poisson point process models, Hard-Sphere models[2], and Strauss Process models. In this paper, we explore an empirical choice, and also provide some comparison with a Poisson point process. We favour the empirical model over the other models due to its ability to better match the PCF shown in Figure 4(a).

Empirical choice for potentials

The choice of $\phi(\boldsymbol{r})$ and $v(\boldsymbol{r}_1, \boldsymbol{r}_2)$ has a profound influence on the resulting model. Since, to the authors' knowledge, this work is the first to propose a PIM-type model for bacteria spatial positions, we have little intuition about the forms of ϕ and v, so would like to estimate them from data. Unfortunately, estimating $v(\boldsymbol{r}_1, \boldsymbol{r}_2)$, a function of six coordinates, empirically requires a vast amount of data; beyond what is available to us from experiment. Thus, we assume several conditions to reduce the dimension of the problem. These assumptions will be described in Section 2.3. The result is that the external potential becomes a function only of z, the coordinate that governs distance from the biofilm-substrate and biofilm-fluid interfaces, and the pairwise potential reduces to a function of z_1, z_2, and the planar separation $|\boldsymbol{r}_1 - \boldsymbol{r}_2|_{xy} \equiv \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$). For a nonnegative, measurable function, $q(\boldsymbol{X})$ of finite point configurations, the expected value over all realisations of the PIM model, with an arbitrary number of points, can be written in the form

$$\mathbb{E}\,[q(\boldsymbol{X})] = \frac{1}{\Xi} \sum_{n=0}^{\infty} \frac{1}{n!} \underbrace{\int \cdots \int}_{n} q(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n) f^{(n)}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n) d\boldsymbol{r}_1, \ldots, d\boldsymbol{r}_n, \tag{2.2}$$

where the normalisation constant, Ξ is known as the grand-canonical partition function [7,26]. The Markov property of the PIM comes from the assumption of $f^{(n)}(\cdot, \ldots, \cdot)$ being dependent on pairwise interactions between particles (or more generally on a Markov function as shown through the Clifford–Hammersley theorem [33,40]).

Poisson point processes

Point processes are point processes with $v(\boldsymbol{r}_1, \boldsymbol{r}_2) = 0$, in other words, there are no interactions between pairs of points. An important property of a Poisson process is that given two disjoint Borel sets, B_1 and B_2, the cardinalities of $\boldsymbol{X}_{B_1}$ and $\boldsymbol{X}_{B_2}$ are independent, Poisson distributed random variables. Numerous characterisations of Poisson point processes exist, and they are frequently used as models for a wide range of random phenomena. The statistical properties of a Poisson point process do not agree with statistical properties of biofilm data because Poisson point processes do not exhibit any interactions between points. However, we will see in Section 6 that mechanical properties

[2] Hard-Sphere processes are point processes that have a minimum radius whereby the set of point configurations with a pair of points closer than the minimum radius is observed with probability 0.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

of data generated from a Poisson process with an average number density equal to that of the biofilm data do match experimental values.

In the next three subsections, we discuss an important class of functions that arise in the study of point processes, show how the external potential and pairwise potential can be estimated from readily observable properties of the data, and provide more detail on the assumptions introduced to reduce the dimensionality of the PIM potentials.

2.1 Factorial moment measures, product densities, and the pair correlation function

The theory of point processes [12, 33] provides a starting point from which to develop a model of the locations of bacteria in biofilms. Point processes are defined as measurable mappings from a probability space, to the space of locally finite point configurations. A locally finite point configuration, $\boldsymbol{X} \equiv \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \}$, is a countable collection of distinct points in $\mathbb{R}^3$ such that for all bounded Borel sets, $B \subset \mathbb{R}^3$, the restriction,$\boldsymbol{X}_B \equiv \boldsymbol{X} \cap B$ is of finite cardinality. Point processes are typically characterised in terms of their probability distribution over locally finite point configurations [33, 45, Section 2] or as locally finite random counting measures [12, Section 9].

In the analysis of point processes, a set of measures known as the *factorial moment measures* (FMM) are of fundamental importance. For a general point process, under mild restrictions (see [12, Section 9.2]), the FMMs can be written in terms of a density, known as a product density, with respect to the Lebesgue measure. Given a point process Φ, the first-order FMM is defined on a Borel set, B as

$$\alpha^{(1)}(B) = \mathbb{E}\left[\sum_{\boldsymbol{x} \in X} \mathbf{1}\left(\boldsymbol{x} \in B\right)\right] = \int \mathbf{1}(\boldsymbol{r} \in B)\rho^{(1)}(\boldsymbol{r})d\boldsymbol{r},$$

i.e., the expectation, over all locally finite point configurations, $\boldsymbol{X}$ distributed according to Φ (written $\boldsymbol{X} \sim \Phi$ henceforth), of the cardinality of $\boldsymbol{X}_B$. The second-order FMM is defined by

$$\begin{aligned}\alpha^{(2)}(B_1 \times B_2) &= \mathbb{E}\left[\sum_{\boldsymbol{x} \in X} \sum_{\boldsymbol{y} \in X \setminus \boldsymbol{x}} \mathbf{1}\left(\boldsymbol{x} \in B_1\right)\mathbf{1}\left(\boldsymbol{y} \in B_2\right)\right] \\ &= \int \int \mathbf{1}(\boldsymbol{r}_1 \in B_1)\mathbf{1}(\boldsymbol{r}_2 \in B_2)\rho^{(2)}(\boldsymbol{r}_1, \boldsymbol{r}_2)d\boldsymbol{r}_1 d\boldsymbol{r}_2.\end{aligned}$$

The first-order product density, $\rho^{(1)}$, is called the *number density* or *intensity*. In this work, we favour the term *number density* over *intensity* since the raw experimental data has a greyscale intensity that does not correspond to the definition of intensity typically used in spatial point process literature. The density function, $\rho^{(2)}$ is called the *second-order product density*.

The PCF, defined as

$$g(\boldsymbol{r}_1, \boldsymbol{r}_2) = \frac{\rho^{(2)}(\boldsymbol{r}_1, \boldsymbol{r}_2)}{\rho^{(1)}(\boldsymbol{r}_1)\rho^{(1)}(\boldsymbol{r}_2)},$$

is also of great interest. It provides a characterisation of the likelihood of two points $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ occurring simultaneously in a realisation of Φ relative to a Poisson process with

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

number density $\rho^{(1)}(\boldsymbol{r})$. Estimates of $g(\boldsymbol{r}_1, \boldsymbol{r}_2)$ are needed in Sections 4 and 6 to compare the PIM model to solve certain integral equations and to compare the biofilm model to experimental data.

For PIMs, the FMMs and product densities can be related through functional differentiation techniques to Ξ, the grand canonical partition function.

2.2 Integral equations for the external and pair potentials

With pairwise interactions, the relation between $\rho^{(1)}$ and ϕ is complicated; thus, it is convenient to define a new function, $a(\boldsymbol{r}) \equiv \exp(-\phi(\boldsymbol{r}))$ (known as the *activity*, i.e., $z(\boldsymbol{r})$ in [26]). From data, we can immediately estimate $\rho^{(1)}$ and g, but not ϕ and v. Thus, we review how integral equations relating ϕ and v to $\rho^{(1)}$ and g are derived through functional differentiation techniques. The discussion here summarises [26, Section 5], and numerical approximations methods used to obtain ϕ and v from the data are described in Section 4.

Given an external potential and using the definition of the functional derivative[3],

$$\phi(\boldsymbol{r}_1 + \boldsymbol{s}) = \phi(\boldsymbol{s}) + \int \frac{\delta\phi(\boldsymbol{r}_1)}{\delta\rho^{(1)}(\boldsymbol{r}_2)} \left(\rho^{(1)}(\boldsymbol{r}_2 + \boldsymbol{s}) - \rho^{(1)}(\boldsymbol{s})\right) d\boldsymbol{r}_2. \tag{2.3}$$

To obtain $\delta\phi/\delta\rho^{(1)}$, it ends up being convenient to define a related function,

$$c(\boldsymbol{r}_1, \boldsymbol{r}_2) \equiv \frac{\delta \log\left(a(\boldsymbol{r}_1)/\rho^{(1)}(\boldsymbol{r}_1)\right)}{\delta\rho^{(1)}(\boldsymbol{r}_2)},$$

known as the *direct correlation function* (DCF). From the definition of $\phi(\boldsymbol{r}_1)$,

$$\frac{\delta\phi(\boldsymbol{r}_1)}{\delta\rho^{(1)}(\boldsymbol{r}_2)} = -\frac{\delta \log a(\boldsymbol{r}_1)}{\delta\rho^{(1)}(\boldsymbol{r}_2)} = -\frac{1}{\rho^{(1)}(\boldsymbol{r}_1)}\delta(\boldsymbol{r}_1 - \boldsymbol{r}_2) + c(\boldsymbol{r}_1, \boldsymbol{r}_2). \tag{2.4}$$

Using this result in equation (2.3), and taking the limit as $\boldsymbol{s} \to \boldsymbol{0}$, an equation for ϕ is obtained,

$$\nabla \log \rho^{(1)}(\boldsymbol{r}_1) + \nabla\phi(\boldsymbol{r}_1) = \int c(\boldsymbol{r}_1, \boldsymbol{r}_2)\nabla_{\boldsymbol{r}_2}\rho^{(1)}(\boldsymbol{r}_2)d\boldsymbol{r}_2. \tag{2.5}$$

The integration is over all of space, thus, includes any regions, where $\rho^{(1)}$ may exhibit a discontinuity due to domain boundaries [32]. Noting that the derivative of a jump discontinuity in the distributional sense leads to a Dirac delta, we may expand equation (2.5) as

$$\nabla \log \rho^{(1)}(\boldsymbol{r}_1) + \nabla\phi(\boldsymbol{r}_1) = \int_V c(\boldsymbol{r}_1, \boldsymbol{r}_2)\nabla_{\boldsymbol{r}_2}\rho^{(1)}(\boldsymbol{r}_2)d\boldsymbol{r}_2 + \int_{\partial V} c(\boldsymbol{r}_1, \boldsymbol{r}_2)\rho^{(1)}(\boldsymbol{r}_2)d\boldsymbol{r}_2,$$

where V is the domain where $\rho^{(1)}$ is positive, and ∂V is the boundary of V. As noted in [32], it is also possible to integrate by parts to eliminate the boundary terms; however,

[3] Given a functional, $\Psi(f)$, the functional derivative [9, ch. 4 Section 3] with respect to $f(\boldsymbol{r})$ is defined by the relation

$$\int \frac{\delta\Psi}{\delta f}(\boldsymbol{r})q(\boldsymbol{r})d\boldsymbol{r} = \frac{d}{d\epsilon}\left[\Psi(f(\boldsymbol{r}) + \epsilon q(\boldsymbol{r}))\right]_{\epsilon=0}.$$

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

this requires approximation of the gradient of $c(\boldsymbol{r}_1, \boldsymbol{r}_2)$. However, such an approach is more difficult numerically than approximating the number density derivative.

Now, we must derive an expression for $c(\boldsymbol{r}_1, \boldsymbol{r}_2)$ in terms of computable quantities. To solve for $c(\boldsymbol{r}_1, \boldsymbol{r}_2)$, we use the functional differentiation identity

$$\int \frac{\delta\phi(\boldsymbol{r}_1)}{\delta\rho^{(1)}(\boldsymbol{r}_3)} \frac{\delta\rho^{(1)}(\boldsymbol{r}_3)}{\delta\phi(\boldsymbol{r}_2)} d\boldsymbol{r}' = \delta(\boldsymbol{r}_1 - \boldsymbol{r}_2). \tag{2.6}$$

Defining $h(\boldsymbol{r}_1, \boldsymbol{r}_2) \equiv g(\boldsymbol{r}_1, \boldsymbol{r}_2) - 1$, it is possible to show that

$$\begin{aligned} \frac{\delta\rho^{(1)}(\boldsymbol{r}_1)}{\delta\phi(\boldsymbol{r}_2)} = \frac{\delta^2 \log \Xi}{\delta a(\boldsymbol{r}_1)\delta a(\boldsymbol{r}_2)} &= \rho^{(1)}(\boldsymbol{r}_1)\rho^{(1)}(\boldsymbol{r}_2) - \delta(\boldsymbol{r}_1 - \boldsymbol{r}_2)\rho^{(1)}(\boldsymbol{r}_1) - \rho^{(2)}(\boldsymbol{r}_1, \boldsymbol{r}_2). \\ &= \delta(\boldsymbol{r}_1 - \boldsymbol{r}_2)\rho^{(1)}(\boldsymbol{r}_1) - h(\boldsymbol{r}_1, \boldsymbol{r}_2). \end{aligned} \tag{2.7}$$

Substituting our results from equations (2.7) and (2.4) into relation (2.6), we obtain an equation known as the Ornstein–Zernike equation [34]

$$h(\boldsymbol{r}_1, \boldsymbol{r}_2) = c(\boldsymbol{r}_1, \boldsymbol{r}_2) + \int \rho^{(1)}(\boldsymbol{r}_3) c(\boldsymbol{r}_1, \boldsymbol{r}_3) h(\boldsymbol{r}_3, \boldsymbol{r}_2) d\boldsymbol{r}_3. \tag{2.8}$$

This equation provides a definition for $c(\boldsymbol{r}_1, \boldsymbol{r}_2)$ in terms of known quantities, and thus an equation for the density variation can be solved once $c(\boldsymbol{r}_1, \boldsymbol{r}_2)$ is found.

Next, we seek a relation between $c(\boldsymbol{r}_1, \boldsymbol{r}_2)$, $h(\boldsymbol{r}_1, \boldsymbol{r}_2)$, and $v(\boldsymbol{r}_1, \boldsymbol{r}_2)$. This can be done using the functional expansion technique described in [26, Section 5]. The strategy is to assume a pairwise interaction system with a particle fixed at $\boldsymbol{x} = \boldsymbol{R}$. This particle induces a potential of the form

$$\phi(\boldsymbol{r}) = v(\boldsymbol{r}, \boldsymbol{R}).$$

Expanding the function $\log(\rho_\phi^{(1)}(\boldsymbol{r})/a_\phi(\boldsymbol{r}))$ in terms of $\rho_\phi^{(1)}(\boldsymbol{r})$ about the reference number density $\rho_0^{(1)}(\boldsymbol{r})$ (i.e., the number density with $\phi(\boldsymbol{r}) = 0$), we obtain

$$\begin{aligned} \log\left(\frac{\rho_\phi^{(1)}(\boldsymbol{r})}{a_\phi(\boldsymbol{r})}\right) = \log\left(\frac{\rho_0^{(1)}(\boldsymbol{r})}{a_0(\boldsymbol{r})}\right) &+ \int \left[\frac{\delta}{\delta\rho^{(1)}(\boldsymbol{r}')} \log\left(\frac{\rho_\phi^{(1)}(\boldsymbol{r})}{a_\phi(\boldsymbol{r})}\right)\right]_{\phi=0} \left(\rho_\phi^{(1)}(\boldsymbol{r}') - \rho_0^{(1)}(\boldsymbol{r}')\right) d\boldsymbol{r}' \\ &+ \mathcal{O}\left(\left|\rho_\phi^{(1)}(\boldsymbol{r}) - \rho_0^{(1)}(\boldsymbol{r})\right|^2\right). \end{aligned}$$

By noting that $\rho_\phi^{(1)}(\boldsymbol{r}) = \rho_0^{(1)}(\boldsymbol{r}) g_0(\boldsymbol{r}, \boldsymbol{R})$ and $a_\phi(\boldsymbol{r}) = a_0(\boldsymbol{r}) \exp(-\phi(\boldsymbol{r}))$, the left-hand side reduces to $\log g_0(\boldsymbol{r}, \boldsymbol{R}) + v(\boldsymbol{r}, \boldsymbol{R}) + \log\left(\rho_0^{(1)}(\boldsymbol{r})/a_0(\boldsymbol{r})\right)$, and by ignoring higher-order terms, the right-hand side is simplified using equation (2.8) to yield

$$v(\boldsymbol{r}, \boldsymbol{R}) = h(\boldsymbol{r}, \boldsymbol{R}) - c(\boldsymbol{r}, \boldsymbol{R}) - \log g(\boldsymbol{r}, \boldsymbol{R}). \tag{2.9}$$

This result is known as the HNC [26]. It relates the pair potential to the DCF and PCF.

It can be seen that the external potential and pair correlation can be found by first

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

solving the the Ornstein–Zernike equation for $c(\boldsymbol{r}_1, \boldsymbol{r}_2)$, then computing $\phi(\boldsymbol{r})$ from equation (2.5), and finally, solving for $v(\boldsymbol{r}_1, \boldsymbol{r}_2)$ in terms of $c(\boldsymbol{r}_1, \boldsymbol{r}_2)$ and $g(\boldsymbol{r}_1, \boldsymbol{r}_2)$ through equation (2.9).

2.3 Simplifying assumptions

At the beginning of this Section 2.1, we mentioned that certain assumptions are made to reduce the dimensionality of the integral equations. We now describe several concepts that lead to the assumptions we make.

In [4], the concept of a *second-order intensity reweighted stationary (* SOIRS) point process was introduced. They are defined for strictly positive number densities in terms of the measure

$$M(B_1 \times B_2) = \mathbb{E}\left[\sum_{\boldsymbol{x} \in \boldsymbol{X} \cap B_1} \sum_{\boldsymbol{y} \in \boldsymbol{X} \cap B_2} \frac{1}{\rho(\boldsymbol{x})} \frac{1}{\rho(\boldsymbol{y})}\right].$$

This measure is closely related to the second-order factorial moment measure, $\alpha^{(2)}(B_1 \times B_2)$. In particular,

$$M(B_1 \times B_2) = \int_{B_1} \int_{B_2} \frac{\rho^{(2)}(\boldsymbol{r}_1, \boldsymbol{r}_2)}{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)} d\boldsymbol{r}_1 d\boldsymbol{r}_2 = \int_{B_1} \int_{B_2} g(\boldsymbol{r}_1, \boldsymbol{r}_2) d\boldsymbol{r}_1 d\boldsymbol{r}_2.$$

If $M(B_1 \times B_2)$ is translation invariant, we say that the point process is SOIRS. This implies that the PCF of a SOIRS point process is translation invariant. Thus, $g(\boldsymbol{r}_1, \boldsymbol{r}_2) = g(\boldsymbol{r}_1 - \boldsymbol{r}_2)$. Our first assumption is that the biofilm data results from realisations of a SOIRS point process.

The SOIRS assumption leads to a PCF dependent on a single 3-D variable representing the vectorial difference between two points. However, the further assumption of isotropy, or that only the magnitude of the distance, but not the angular orientation between the pair of points matters, allows for the simplification, $g(\boldsymbol{r}_1 - \boldsymbol{r}_2) = g(|\boldsymbol{r}_1 - \boldsymbol{r}_2|)$.

Second, we assume that the density varies only as a function of distance from the substrate. We make this assumption due to symmetry considerations; the bacteria are grown on a uniformly planar plate. Thus, we do not expect any difference in the number density for any two points in space that are of equal distance from the plate.

We show in appendix that a number density that varies along one dimension, and an isotropic, translation invariant PCF are consistent with a DCF (and pair potential) of the form $c(\boldsymbol{r}_1, \boldsymbol{r}_2) = c(|\boldsymbol{r}_1 - \boldsymbol{r}_2|, z_1, z_2)$. Defining $r_{xy} \equiv |\boldsymbol{r}|_{xy} = \sqrt{x^2 + y^2}$, equation (2.5) reduces to

$$\begin{aligned} \frac{d}{dz_1}(\log \rho(z_1) + \phi(z_1)) = 2\pi \int_{z_0}^{z_{\max}} \left(\int_0^\infty c(r_{xy}, z_1, z_2) r_{xy} dr_{xy}\right) \frac{d}{dz_2} \rho(z_2) dz_2 \\ + 2\pi \left[\rho(0) \int_0^\infty c(r_{xy}, z_1, 0) r_{xy} dr_{xy}\right. \\ \left. - \rho(z_{\max}) \int_0^\infty c(r_{xy}, z_1, z_{\max}) r_{xy} dr_{xy}\right]. \end{aligned} \tag{2.10}$$

Formulas of this type appear to have first been derived in the context of understanding

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

fluid interfaces in [32]. Applying a transformation $\boldsymbol{r}_3' = \boldsymbol{r}_3 - \boldsymbol{r}_2$, equation (2.8) reduces to

$$h(|\boldsymbol{r}_1 - \boldsymbol{r}_2|) = c(|\boldsymbol{r}_1 - \boldsymbol{r}_2|_{xy}, z_1, z_2) + \iiint \rho(z_3' + z_2) h(|\boldsymbol{r}_3'|) c(|\boldsymbol{r}_1 - \boldsymbol{r}_2 - \boldsymbol{r}_3'|_{xy}, z_1, z_3' + z_2) dx_3' dy_3' dz_3'. \quad (2.11)$$

Similarly, the HNC (equation (2.9)) becomes

$$v(|\boldsymbol{r}_1 - \boldsymbol{r}_2|_{xy}, z_1, z_2) = h(|r_1 - \boldsymbol{r}_2|) - c(|\boldsymbol{r}_1 - \boldsymbol{r}_2|_{xy}, z_1, z_2) - \log g(|\boldsymbol{r}_1 - \boldsymbol{r}_2|).$$

With these assumptions, the nth-order probability densities associated with the PIM are of the form

$$f^{(n)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{1}{Z_n} \exp\left(- \sum_{i=1}^{n} \phi(z_i) - \sum_{i=1}^{n} \sum_{j>i}^{n} v(|\boldsymbol{x}_i - \boldsymbol{x}_j|, z_i, z_j) \right).$$

To summarise, the assumptions are listed below

(1) The process is a SOIRS point process [4, 33].

(2) The PCF is isotropic.

(3) The number density is only dependent on the distance from the fluid-biofilm interface, i.e., $\rho(z) \equiv \rho^{(1)}(\boldsymbol{r} = (x, y, z))$.

It remains to (i) estimate the number density and PCF, (ii) numerically solve the OZ equation and density integral equation, and (iii) justify our model assumptions. These are the subjects of the next two sections.

3 Estimation of summary statistics

In this section, we discuss the estimation of $\rho(z)$ and $g(r)$, and the nearest neighbour distributions, defined in Section 3.3, denoted $H_k(r)$. In the case of empirical data, without prior intuition about the point process, non-parametric estimators are often the tool of choice for the estimation of these characteristics. This is the approach we take since we are not aware of any previous models of the spatial statistics of bacteria in biofilms. For notation, we use a hat (i.e., $\hat{\rho}$) to denote an estimator for a quantity.

In addition to being flexible, non-parametric estimates are often easy to implement and versatile, yielding suitable intensity approximations for many classes of point processes. However, a common source of concern with non-parametric estimates is the presence of statistical bias [24, 38, 51]. For ρ, g, and H_k, non-parametric estimates are rarely pointwise unbiased [38], but are often asymptotically unbiased. Heuristically, the bias is due to inaccuracy in the numerical approximation of a Radon–Nikodym derivative [5] and is analogous to the truncation error associated with finite difference approximations of differentiable functions. Furthermore, as finite difference approximations depend on a spatial increment, non-parametric number density estimators are parametrised[4] by a

[4] The term *nonparametric* refers to the absence of an underlying assumption about the functional form of $\rho(\boldsymbol{r})$, not the absence of any tuning parameters in the estimator. Aside from a few simple

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

bandwidth, b. Smaller values of b reduce the bias, but increase the variance of the estimator [35, 38, 51]. To ensure accurate estimators, careful selection of b is needed to balance this tradeoff between bias and variance. Further discussion of this balancing of bias and variance is included in the appendix.

3.1 Number density estimators

As discussed in Section 2.1, the number density of $\boldsymbol{X} \sim \Phi$ is defined through the relation

$$\mathbb{E}\left[\sum_{\boldsymbol{x}\in\boldsymbol{X}} 1(\boldsymbol{x}\in B)\right] = \int_B \rho^{(1)}(\boldsymbol{r})d\boldsymbol{r}.$$

Thus, we see that $\sum_{\boldsymbol{x}\in\boldsymbol{X}} \mathbf{1}(\boldsymbol{x}\in B)$ is an unbiased estimator for $\alpha^{(1)}(B)$. However, $\rho^{(1)}(\boldsymbol{r})$ cannot generally be estimated without bias, since it is defined as the average over an infinitesimal region

$$\rho^{(1)}(\boldsymbol{r}) \equiv \lim_{r\to 0} \frac{\alpha^{(1)}(B_r(\boldsymbol{r}))}{v(B_r(\boldsymbol{r}))},$$

where $v(\cdot)$ is the Lebesgue measure and B_r is a ball of radius r centred at $\boldsymbol{r}$. We see that this is also related to the Dirac delta distribution through the relation

$$\rho^{(1)}(\boldsymbol{r}) = \lim_{r\to 0}\int_B \frac{1}{4\pi r^3} 1(|\boldsymbol{r}'-\boldsymbol{r}|\leqslant r)\rho^{(1)}(\boldsymbol{r}')d\boldsymbol{r}' = \int \delta(\boldsymbol{r}-\boldsymbol{r}')\rho^{(1)}(\boldsymbol{r}')d\boldsymbol{r}'.$$

This motivates an approximation of the form

$$\rho^{(1)}(\boldsymbol{r}) \approx \int k_b(\boldsymbol{r}-\boldsymbol{r}')\rho^{(1)}(\boldsymbol{r}')d\boldsymbol{r}' = \mathbb{E}\left[\sum_{\boldsymbol{x}\in\boldsymbol{X}} k_b(\boldsymbol{r}-\boldsymbol{x})\right]$$

and an estimator

$$\hat{\rho}(\boldsymbol{r}) = \sum_{\boldsymbol{x}\in\boldsymbol{X}} k_b(\boldsymbol{r}-\boldsymbol{x}),$$

where $k_b(\cdot)$ is known as a kernel density and b is the bandwidth. For data in $\mathbb{R}^n$, $k_b(\cdot)$ is typically a radially symmetric function of the form $k_b(\cdot) = \frac{1}{b^n}k_1(\cdot/b)$. There exist many choices for $k_b(\cdot)$; however, as long as $\int k_1(\boldsymbol{r})d\boldsymbol{r} = 1$ and $k_1(\cdot)$ is compactly supported, the resulting estimators are typically relatively insensitive to the choice of $k_1(\cdot)$, but highly sensitive to the choice of b. Last, because of the distributional convergence of $\int k_b(\boldsymbol{r}-\boldsymbol{r}')\phi(\boldsymbol{r}')d\boldsymbol{r}'$ to $\int \delta(\boldsymbol{r}-\boldsymbol{r}')\phi(\boldsymbol{r}')d\boldsymbol{r}'$ in the limit as $b\to 0$, we say that $\hat{\rho}(z)$ is asymptotically unbiased. This estimator is similar to the number density estimator discussed in [45, Section 4]. For what follows, we specify $k_b(z)$ as

$$k_b(z) = \left(\frac{3}{4b}\right)\left(1-\left(z/b\right)^2\right)\mathbf{1}(|z|<b).$$

This kernel is known as the *Epanechnikov kernel* [14], a commonly used density estimation kernel.

examples, non-parametric estimators almost always have some sort of bandwidth parameter that must be chosen based on the length scales over which variations are seen in the data.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

Due to the limited size of a data set, we consider the dependence of the intensity on x-, y-, and z-coordinates separately (for instance, assume that $\rho(\boldsymbol{r})$ only depends on x, independent of y and z). Given the restriction $\boldsymbol{X}_W$, this leads to estimators of the form

$$\hat{\rho}(z) = \sum_{\boldsymbol{x}'=(x',y',z')\in \boldsymbol{X}_W} \frac{k_b(z-z')}{c_{b,W}(z)}. \tag{3.1}$$

Defining A as the area of a cross-section perpendicular to the z-axis and W to be a set in $\mathbb{R}^3$ containing the data, the denominator, $c_{b,W}(z)$ in equation (3.1), is an edge correction factor defined as

$$c_{b,W}(z) = A\int_{z-b}^{\min(z+b,z_{\max})} k_b(z'-z)dz', \tag{3.2}$$

where $z_{\max}$ is the largest value of the z-coordinate of points in W. We see that for $z_{\max}-z<b$ given a differentiable number density

$$\begin{aligned} &A\int_{z-b}^{z_{\max}} k_b(z-z')\rho^{(1)}(z')dz' - A\rho^{(1)}(z) \\ &\quad = A\rho^{(1)}(z)\left(\int_{z-b}^{z_{\max}} k_b(z-z')dz' - 1\right) + \mathcal{O}\left(Ab\frac{d}{dz}\rho^{(1)}(z)\right). \end{aligned}$$

Thus, if we choose $c_{b,W}(z)$ as in equation (3.2), the dominant, zeroth-order error term is eliminated.

Upon computation, we found that although the bacteria locations are 3-D coordinates, the number density only varies significantly along the z-axis, with changes in the distance from the fluid-biofilm interface. This justifies Assumption 1 of Section 2.3. As depicted in Figure 2(a), it appears as though bacteria near the fluid-biofilm interface stratify into layers and pack more closely together than cells further interior. In Figure 2(b), this phenomenon is clearly seen as bumps in the number density.

The number density estimates obtained from data sets #1–4 exhibit some variation near the biofilm-fluid interface; thus, in each simulation of Section 6, we parametrise our model on the number density from one of the data sets (e.g., from data sets #1–4). In Section 6, we separate the comparisons between model data and experimental data by data set. (see Figure 8(a) and (b)).

We also found that the precise form of the number density variation near the top of the biofilm did not have a substantial impact on the simulations conducted in Section 6. The sensitivity of those simulations to variability in the number density was examined by solving the integral equations of Section 2 with a constant number density (set to the number of points in W divided by $\mathrm{Vol}(W)$), and then conducting simulations equivalent to those in Section 6. We found that the dynamic moduli (discussed in Section 6) were very close to those of the variable density data. We expect that with more extreme variation in the number density, the similarity would disappear.

As previously mentioned, the value of the bandwidth, b, has a significant impact on the resulting number density estimate and should be carefully chosen to balance variance and bias. If b is too small, the estimate will be noisy, whereas if b is too large, key features of the data will be blurred. The estimation of a one-dimensional (1-D) intensity function is quite similar to the non-parametric estimation of a probability density up to the

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

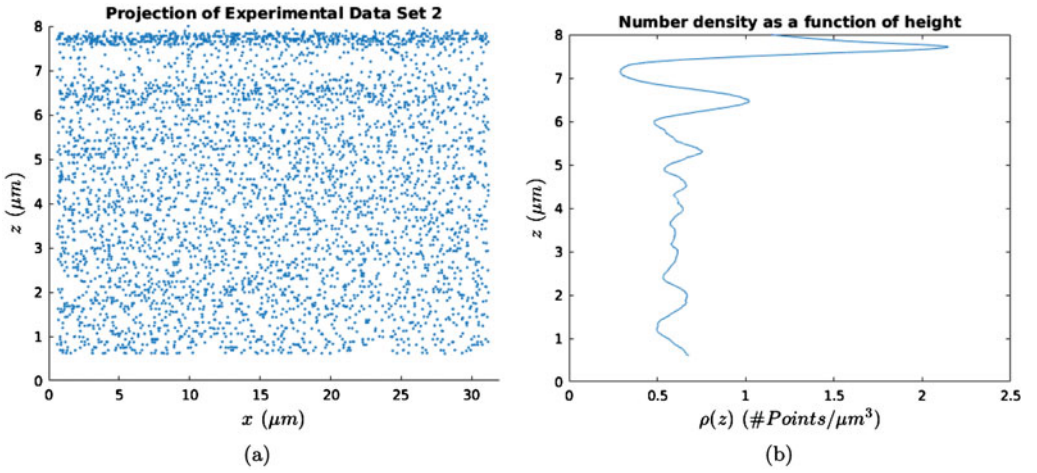

FIGURE 2. (a) The projection of one of the data sets into the xz-plane is shown. The number density variation near the top of the domain can clearly be observed and some oscillatory intensity variation is discernible. (b) The height dependent intensity, $\rho(y)$, from the same experimental data set is depicted. We used a bandwidth, $b = 0.2$ for all four data sets.

normalisation condition required of probability densities. This motivates the adaptation of the *Least Squares Cross-Validation* (LSCV) technique, a common optimisation strategy for non-parametric probability density estimators [51], to optimise our choice for b in equation (3.1). We found that although the LSCV gave a reasonable estimate of b, the result could be improved upon by adapting the Bayesian technique in [18] with a prior centred about the value of b estimated by LSCV. We refer the interested reader to appendix for a discussion of LSCV bandwidth selection and the Bayesian approach of [18].

In addition to the number density, equation (2.5) depends on the gradient of the number density. In order to compute this quantity, we adhere to the strategy discussed in [51] and use the differentiable, triweight kernel density estimator, defined as

$$T_b(z) = \frac{35}{32b}(1 - (z/b)^2)^3 \mathbf{1}_{\{|z| \leqslant b\}}.$$

Similar to the Epanechnikov kernel, the triweight kernel is a symmetric probability density function. However, it is more useful for density derivative estimation since it is twice differentiable, whereas the Epanechnikov kernel is not differentiable. The number density derivative, ρ', is approximated by

$$\hat{\rho}'(z) = \frac{1}{A} \sum_{\boldsymbol{x}' \in \boldsymbol{X}_W} \frac{\partial T_b(z - z')}{\partial z}.$$

As derivatives are more sensitive to noisy data, optimal balancing of variance and bias in number density derivative estimation tends to yield a larger value of b than that used in number density estimation [27]. By inspection of the resulting estimators, we found that a value of $b = 0.5$ seemed to give the best result. In particular, at $b = 0.5$, we found that $\hat{\rho}'$ was not overly noisy, and agreed well with finite difference estimates of the derivative

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

of $\hat{\rho}$. However, in contrast to a finite difference estimate, it is very straightforward to compute $\hat{\rho}'$ for any value of z. To correct for bias near the boundaries of the domain, adjustments to the kernel constructed by the technique in [28, equation (8.2)] were used.

3.2 Pair correlation function estimates

Since the PCF is assumed to be stationary, and isotropic, $g(\boldsymbol{r}_1, \boldsymbol{r}_2)$, generally a function of six coordinates, is simplified to become $g(|\boldsymbol{r}_1 - \boldsymbol{r}_2|)$, a function of one coordinate. The advantages of this simplification for non-parametric estimation is substantial. Following [4, 20], an asymptotically unbiased estimator of the stationary PCF for a SOIRS point process with known $\rho(\boldsymbol{r})$ observed in a set, $W \subset \mathbb{R}^3$ is

$$\hat{g}_1(r) = \frac{1}{4\pi r^2 \bar{\gamma}_W(r)} \sum_{\boldsymbol{x} \in \boldsymbol{X}_W} \sum_{\boldsymbol{y} \in \boldsymbol{X}_W \setminus \boldsymbol{x}} \frac{k_b(r - |\boldsymbol{x} - \boldsymbol{y}|)}{\rho(\boldsymbol{x} \cdot \hat{\boldsymbol{e}}_z)\rho(\boldsymbol{y} \cdot \hat{\boldsymbol{e}}_z)}, \tag{3.3}$$

where $\bar{\gamma}_W(r)$ is the isotropised set covariance of W [45], defined by

$$\bar{\gamma}_W(r) = \int v(W \cap W_{\boldsymbol{t}}) d\boldsymbol{t}$$

with integration occurring over the surface of a sphere of radius r in $\mathbb{R}^3$. We also note that discussions of alternative choices for PCF estimates can be found in [29, 30, 44, 47]. We chose the estimator in equation (3.5), since it yields a continuous estimate of $g(r)$, and because it is relatively easy to implement (although the choice of b presents some difficulty).

Following [44], the expectation of $\hat{g}(r)$ is of the form:

$$\mathbb{E}\,[\hat{g}_1(r)] = \int g(|\boldsymbol{r}'|) k_b(|\boldsymbol{r}'| - r) d\boldsymbol{r}'. \tag{3.4}$$

Taking the limit as $b \to 0$ in equation (3.4) shows that, under a few restraints on $k_b(r)$, $\hat{g}_1(r) \to g(r)$ at every point of continuity of $g(r)$. Thus, $\hat{g}_1(r)$ is asymptotically pointwise unbiased for continuous PCFs[5].

Since $\rho(\boldsymbol{r})$ is not known in our application, we use an estimator of the form

$$\hat{g}(r) = \frac{1}{4\pi r^2 \bar{\gamma}_W(r)} \sum_{\boldsymbol{x} \in \boldsymbol{X}_W} \sum_{\boldsymbol{y} \in \boldsymbol{X}_W \setminus \boldsymbol{x}} \frac{k_b(r - |\boldsymbol{x} - \boldsymbol{y}|)}{\hat{\rho}(\boldsymbol{x} \cdot \hat{\boldsymbol{e}}_z)\hat{\rho}(\boldsymbol{y} \cdot \hat{\boldsymbol{e}}_z)}. \tag{3.5}$$

Unlike $\hat{g}_1(r)$, the use of an approximate number density causes $\hat{g}(r)$ to fail to be asymptotically unbiased.

Equation (3.3) is valid for isotropic SOIRS point processes only. An estimator of the form

$$\hat{g}_a(\boldsymbol{r}) = \frac{1}{4\pi r^2 v(W \cap W_{\boldsymbol{r}})} \sum_{\boldsymbol{x} \in \boldsymbol{X}_W} \sum_{\boldsymbol{y} \in \boldsymbol{X}_W \setminus \boldsymbol{x}} \frac{k_b(|\boldsymbol{r} - (\boldsymbol{x} - \boldsymbol{y})|)}{\hat{\rho}(\boldsymbol{x} \cdot \hat{\boldsymbol{e}}_z)\hat{\rho}(\boldsymbol{y} \cdot \hat{\boldsymbol{e}}_z)}, \tag{3.6}$$

can be used to obtain an anisotropic PCF estimate.

[5] For discontinuous PCFs, such as those arising from hard-sphere processes, the estimate is not asymptotically pointwise unbiased at the point of discontinuity, but is asymptotically convergent to $g(r)$ in the least squares sense.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

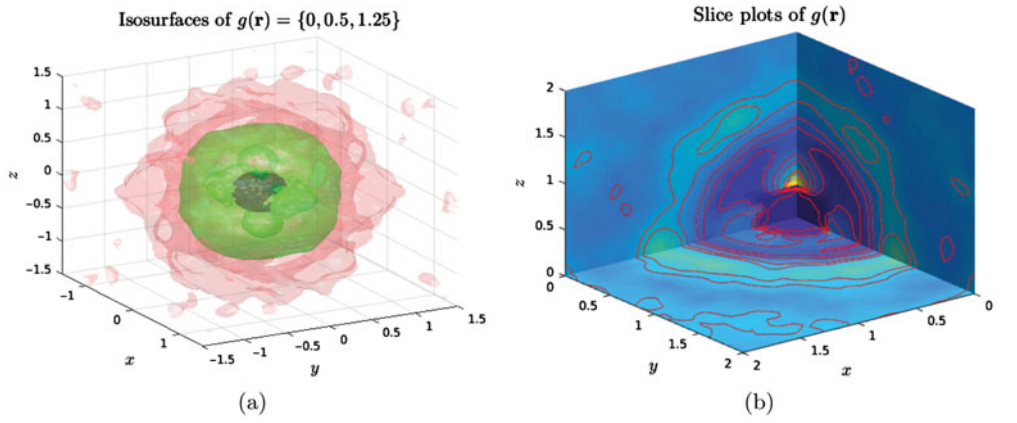

FIGURE 3. Both images are of PCFs averaged over the four experimental data sets. In (a) isosurfaces of the anisotropic pair correlation function at $g^{(2)}(\boldsymbol{r}) = 0$ (inner isosurface), 0.5 (middle), and 1.25 (outer) are depicted, and (b) contains slice plots and contours of the anisotropic pair correlation function. The anisotropic PCF is related to the isotropic PCF through the integral $g(r) = \frac{1}{4\pi} \int \int g_a(r, \phi, \theta) \sin\theta d\theta d\phi$. The anisotropic pair correlation function shows some vertical anisotropy, however, most of the variation appears as a radially symmetric function.

In Figure 3(a) and (b), $\hat{g}_a(\boldsymbol{r})$, averaged over four data sets, is depicted. We display the averaged estimate since the anisotropic estimator is subject to higher variance. Data set averaging is employed as a variance reduction technique.

It can be seen that, although there is some anisotropy, the isosurfaces of $\hat{g}_a(\boldsymbol{r})$ are roughly spherical. Thus, we believe that the isotropic approximation of $g(\boldsymbol{r})$ is reasonable. In light of this result, we will assume in what follows that g is isotropic. In Figure 4, the SOIRS radially symmetric PCF estimator defined in (3.3), computed from one data set, is depicted.

An intriguing property of the biofilm data is the presence of two prominent peaks in the PCF. It has been suggested that the first, smaller peak in Figure 4 is indicative of bacteria undergoing cell division at the time their position was measured [41]. The second peak occurs near the average diameter of a non-dividing bacterium. The third peak at $r = 2.330$ μm is indicative of local regularity in the positions of bacteria [26, Section 5]. Last, we also applied the mode detection test described in [39] to rule out the likelihood of further maxima in the interval $[0, 2.5]$. We found that for the bandwidth's deemed suitable by the LSCV procedure described in appendix, further peaks in the PCF are unlikely.

Recall from Figure 2 that there is evidence of variability in the number density along the z-axis. This motivated our assumption of a SOIRS process in analysing the data. However, we have not provided any evidence against an alternative hypothesis: that the PCF is not translation invariant. To make a convincing argument in favour of the SOIRS assumption (or at least that it is reasonable), it is necessary to estimate the magnitude of the variability of the pair correlation

To test for variability in the PCF, its estimator, $\hat{g}(r)$ is calculated for various subsets of the full data set, $\boldsymbol{X}_W$. In particular, we compute $\hat{g}(r)$ over sets of the form $V_z = [z - \Delta z, z + \Delta z] \cap W$ with $z \in (\min(z_i), \max(z_i))$, where z_i are the z-coordinates of the bacteria, and $\Delta z = 1$ μm. Computing $\hat{g}(r)$ on V_z as a function of z, we found only small

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

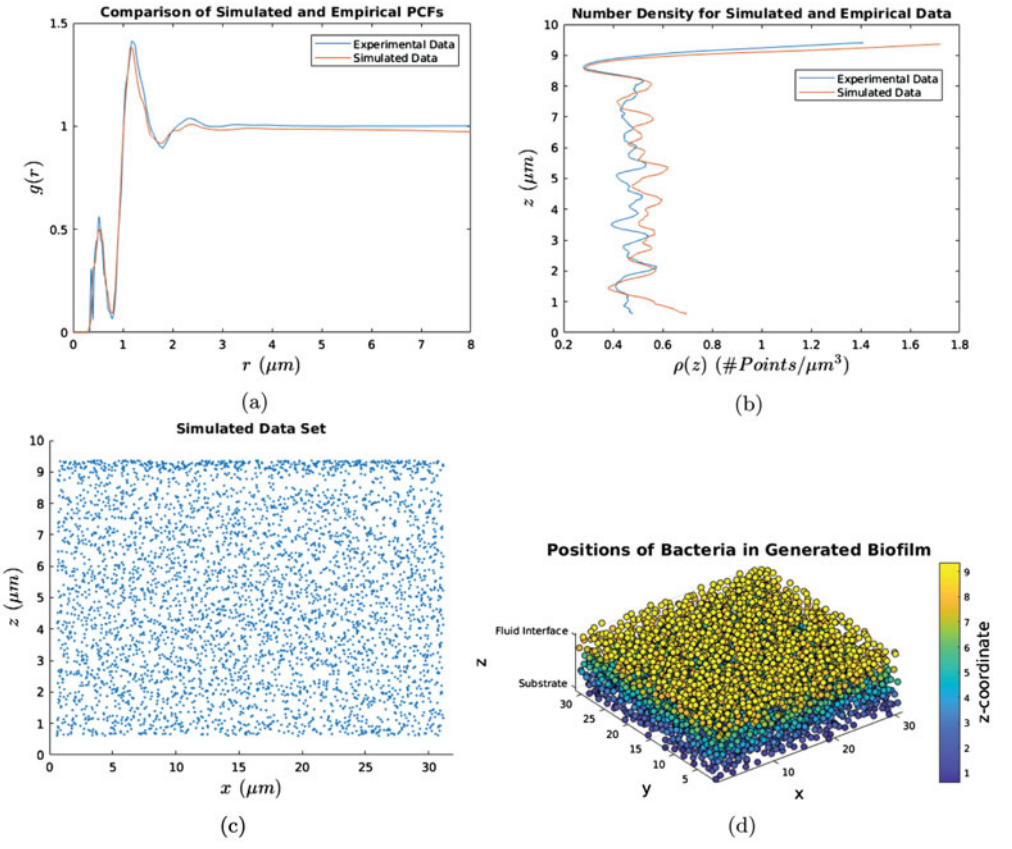

FIGURE 4. (a) Comparison between experimental and simulated data of the SOIRS estimator for a radially symmetric pair correlation function. The radially symmetric pair correlation exhibits peaks at $r = 0.525$, $r = 1.185$, and $r = 2.330$ μm. The relative heights of the first two peaks is an interesting finding, and has been observed in biofilm data in a previous study [42]. The first, smaller peak is likely due to a small number of bacteria that were undergoing cellular division at the time the data set was recorded although the possibility of some experimental error cannot be definitively ruled out [41]. (b) Comparison of number density between experimental data and simulated data. The simulation method used to generate data is discussed in Section 4. (c) Projection of the simulated data set into the xz-plane. (d) Image of a simulated biofilm generated form Algorithm 1.

variations in the resulting PCF. To fully utilise the available data in these computations, we use an altered estimator of $g(r)$, denoted $\hat{g}_{V_z}(r)$

$$\hat{g}_{V_z}(r) = \frac{1}{4\pi r^2 \bar{\gamma}_{V_z,W}(r)} \sum_{\boldsymbol{x} \in \boldsymbol{X}_{V_z}} \sum_{\boldsymbol{y} \in \boldsymbol{X}_W \setminus \boldsymbol{x}} \frac{k_b(r - |\boldsymbol{x} - \boldsymbol{y}|)}{\hat{\rho}(\boldsymbol{x} \cdot \hat{\boldsymbol{e}}_z)\hat{\rho}(\boldsymbol{y} \cdot \hat{\boldsymbol{e}}_z)}$$

with

$$\bar{\gamma}_{V_z,W}(r) = \frac{1}{4\pi r^2} \int v(W \cap V_{z,\boldsymbol{t}}) d\boldsymbol{t}.$$

The integration is carried out over the surface of a sphere of radius r in $\mathbb{R}^3$. This altered estimator is used since it takes into account data that is in W but not V_z in the inner

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

summation. This reduces the truncation errors that could occur if we used the estimator in equation (3.3). It is in essence a generalisation of our standard PCF estimator to allow for the case that $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ correspond to different types of points (or points in different subdomains).

Upon computation, we note that the mean square error [6] over z remained less than 0.05 for $r \in (0, 2)$. We do observe some variation with z in the PCF, however, it seems to be a minor effect. In particular, we note that in some of the data sets, the height of the first peak in $\hat{g}_{V_z}(r)$ decreases from the bottom to the top of W. Although it would be ideal to compute a non-stationary estimate for $g(\boldsymbol{r}_1, \boldsymbol{r}_2)$, in practice, we found our calculations for such an estimator to be noisy, and unreliable for small values of $|\boldsymbol{r}_1 - \boldsymbol{r}_2|$ given the amount of data available. Thus, we have approximated the variable PCF with a SOIRS form of $\hat{g}(r)$ for use in our computations.

As a final note, alternative approaches for estimating $g(r)$ by maximum likelihood estimation and Takacs–Fiksel estimation have been explored in several papers [3,4,33,37]. With maximum likelihood estimation, the pair potential is assumed to be a member of a predetermined class of functions that differ through some set of parameters that can be optimised to fit the data. We did not proceed with this approach due to the unusual structure of the PCF. With Takacs–Fiksel estimation, $g(r)$ is assumed to be a piecewise function (i.e., piecewise polynomial). The weights of the coefficients in the piecewise approximant are found by maximising a non-linear system of equations. Possible issues with such an approach here are the computational cost to accurately resolve $g(r)$ over a range of r and the attainment of a continuous (not just piecewise continuous) result.

3.3 Nearest neighbour distributions

Although not used in the PIM, the nearest neighbour distributions are useful for justifying assumptions made in the development of the PIM. As discussed in Section 2.2, the PIM is parametrised by first- and second-order statistics of the data (through $\rho^{(1)}(z)$ and $g(r)$). Thus, we would expect that the PCF and number density of the PIM will closely match those of the experimental data. However, it remains to be justified that such statistics provide a useful characterisation of the data; it could be the case that higher-order statistics or clustering phenomena are crucial.

It is in general difficult to explicitly demonstrate that many body interactions are negligible since the estimation of higher-order interaction functions, plagued by the curse of dimensionality, requires vast quantities of data. However, the validity of a pairwise interaction assumption may be evidenced by comparisons of lower-order summary statistics that depend on many body interactions. In this capacity, the k-nearest neighbour distributions are useful.

The kth-nearest neighbour distribution is defined as the probability density function of distances between a point $\boldsymbol{x} \in \boldsymbol{X}_W$ and the kth closest point $\boldsymbol{y} \in \boldsymbol{X}_W \backslash \boldsymbol{x}$. Following [48,49], the *nearest neighbour density* (NND), is defined as the conditional probability[7] probability

[6] The mean square error is defined as $E = \int_{r_0}^{r_1} (\hat{g}_{V_z}(r) - \hat{g}(r))^2 dr$. It is approximated by a trapezoid rule quadrature.

[7] For the point processes, we consider, $\forall \boldsymbol{x} \in W$, $\mathbb{P}[\boldsymbol{x} \in \boldsymbol{X}_W] = 0$ since $\boldsymbol{X}_W$ is finite and W, being a continuum, is uncountable. The nearest neighbour distribution is defined rigorously through the

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

over $\boldsymbol{X} \sim \Phi$, given $\boldsymbol{x} \in \boldsymbol{X}$,

$$H_k(r) = \lim_{\Delta r \to 0} \frac{\mathbb{P}\left[\exists\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\} \in \boldsymbol{X} \text{ such that } r \leqslant \max_{[1,\ldots,k]} |\boldsymbol{x} - \boldsymbol{y}_k| \leqslant r + \Delta r | \boldsymbol{x} \in \boldsymbol{X}\right]}{\Delta r}.$$

Along with the intensity and PCF, we use the NNDs as a means of comparing experimental data to realisations of point processes generated through simulation. As shown in [48,49], $H_k(r)$ depends on the product density, $\rho^{(k+1)}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{k+1})$, and not lower-order densities. Although equivalence in NNDs does not guarantee equivalence of two point processes in probability, if two point processes are equal in probability, they must have the same nearest neighbour distributions. In practice, testing for equivalence in probability of a general point process is computationally infeasible, and would require an inordinate amount of data; however, the nearest neighbour distributions are a useful summary statistic due to their low dimensionality, and the ease with which they can be estimated.

Two disadvantages of nearest neighbour distributions as a means of comparing point processes are their lack of directional information, and that, as defined here, they are spatially homogeneous. Thus, the NNDs we compute are best understood as homogenised variants of a more general nearest neighbour distribution that may depend on location. Since only four data sets are available, the computation of a spatially variable NND would be a formidable difficulty, and subject to greater variance than the homogenised NND.

To estimate the nearest neighbour distributions, we use *minus sampling* [45]. Minus sampling is the technique of constructing estimators that 'leave out' points near the edges of the domain to mitigate edge effects. Minus sampling estimators are less efficient than other estimators because they do not utilise all available data. However, for statistics that are strongly influenced by edge effects, the benefit in reducing edge effects can be well worth the inefficient use of data. Using the symbol, $\ominus$, to denote Minkowski subtraction [45, Section 1], the nearest neighbour distribution is estimated as

$$\hat{H}_P(r) = \frac{1}{\Phi\left(W \ominus \mathcal{B}_{r_c}(0)\right)} \sum_{\boldsymbol{x} \in \boldsymbol{X}_W \ominus \mathcal{B}_{r_c}(0)} k_b\left(r - \left|\boldsymbol{x} - \underset{\boldsymbol{y} \in \boldsymbol{X}_W \setminus \boldsymbol{x}}{\arg\min} |\boldsymbol{y} - \boldsymbol{x}|\right|\right).$$

Minus sampling is suitable in our application since the box-shaped geometry of the domain dictates that most of the data is located away from the edges. Thus, the loss of data due to minus sampling is not severe. Even though there is a variation in number density near the top of the domain, we found that computations of the NNDs that left out this portion of the data did not have a pronounced effect on the resulting estimate.

Because $H_k(r)$ is a 1-D probability density, kernel density estimation methods of choosing b based on the cardinality of $\boldsymbol{X}_W$ are used [35, 38]. In particular, we use the *ksdensity* function in Matlab to compute $H_k(r)$ once r_c is known. To specify r_c, we first compute a preliminary estimate, $\hat{H}_{P,0}(r)$, without minus sampling (e.g., assuming $r_c = 0$),

use of Palm distribution theory [7,45]. Roughly speaking, $\mathbb{P}\left[\cdot | \boldsymbol{x} \in \boldsymbol{X}_W\right]$ is the Palm distribution of Φ with respect to the constraint $\boldsymbol{x} \in \boldsymbol{X}_W$.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

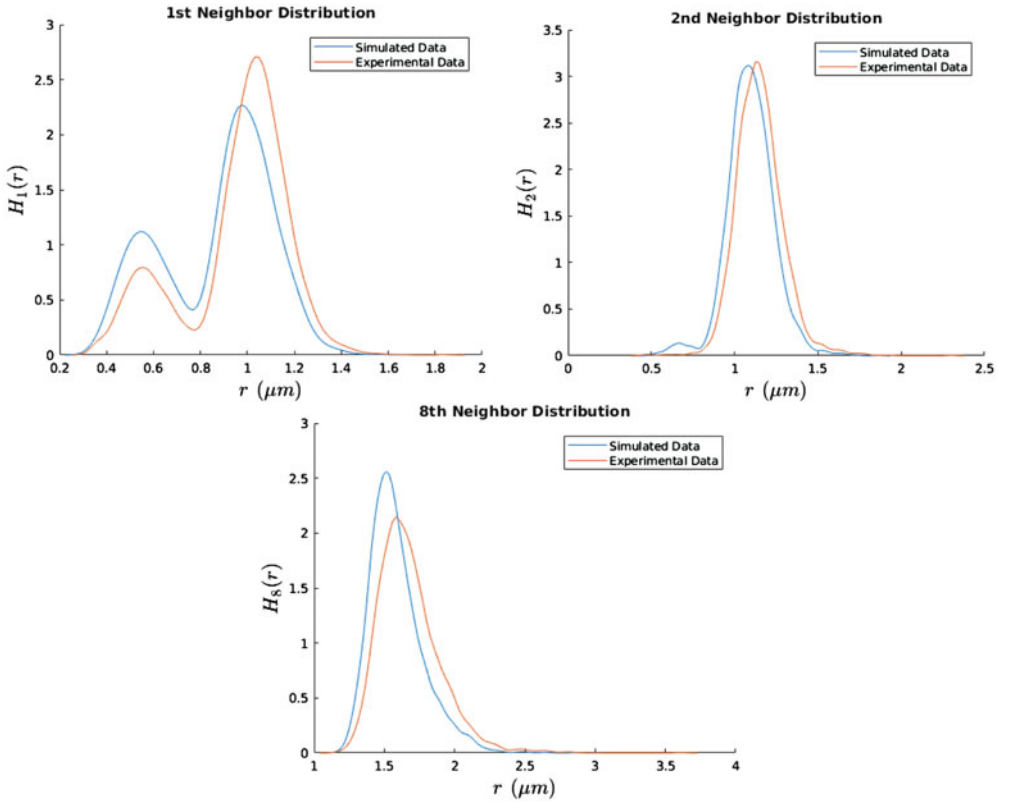

FIGURE 5. Comparison of nearest neighbour density functions between a simulated biofilm, and the experimental data. The bimodal behaviour in the first neighbour distribution is not completely captured by the model; however, higher NNDFs (2 and 8 depicted here) are very accurately matched by the model.

and then take r_c equal to the supremum

$$r_c = \arg\sup_r \left\{ \hat{H}_{P,0}(r) > 0 \right\}.$$

This choice ensures that, with high probability, the nearest neighbour of each point of $\Phi_W \ominus \mathcal{B}_{r_c}(0)$ is contained in Φ_W. Specifically, the expected error is the probability $\Pr[\min_{\boldsymbol{r}_j \in \Phi} |\boldsymbol{r}_i - \boldsymbol{r}_j| > r_c,\ \boldsymbol{r}_i \in \Phi \ominus \mathcal{B}_{r_c}(0)]$. Since each data set contains $\sim 4,000$ bacteria locations, this probability should be on the order of 10^{-3}. A second bias term arises from the use of non-parametric density estimators that have finite support for nonzero b. We expect this source of bias to be small since the asymptotic bias of kernel smoothing is typically $\mathcal{O}(n^{-4/5})$ [35], where n is size of the data set (~4,000 for each biofilm data set). For $k > 1$, analogous estimators are used with r_c replaced by $r_c^{(k)} = \arg\sup_r \left\{ \hat{H}_{P,0}^{(k)}(r) > 0 \right\}$. For larger values of k, the minus sampling technique is expected to become inaccurate as the subset of W from which points may be chosen shrinks. However, we found that $r_c^{(k)}$ does not rapidly increase with k, and at least for $k \leqslant 20$, the amount of data that must be disregarded due to minus sampling is small compared to the amount of data in each data

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

set. In Figure 5, we show the nearest neighbour functions computed for the experimental data (from data set #1) and a realisation of the PIM model of Section (2).

4 Numerical solution of integral equations

After obtaining estimates for $h(|\boldsymbol{r}_1-\boldsymbol{r}_2|)$ and $\rho(z)$ using the methods of Section 3, equations (2.11) and (2.10) must be solved numerically. In the homogeneous case, the integral term in the OZ equation becomes a convolution that can be efficiently and accurately handled by Fourier transform methods. However, due to the inhomogeneous number density, the integral over z is not a convolution in this present application. In fact, with a variable number density, the OZ equation implies that the pair correlation and DCFs cannot both simultaneously be translation invariant. A proof of this fact is straightforward and shown in the appendix. Since the standard Fourier transform methods for convolutional integrals will not apply to our application, we use a two-dimensional (2-D) Hankel transform in the radial direction. To quantify variation in z and z' variation in the DCF, we solve a Fredholm integral equation of the second kind.

As discussed in the Section 2.3, to simplify equation (2.8), we assume that the xy dependence of the pair correlation and DCFs is homogeneous with regard to within-plane translations and is radially symmetric. This assumption makes the xy integration in equation (2.11), a 2-D radially symmetric convolution. The radially symmetric convolution of two functions can be found through use of the 2-D Hankel transform, denoted $\mathcal{H}[\cdot]$, defined as the involutory transform

$$F(k) = \mathcal{H}[f(r)] = 2\pi \int_0^\infty J_0(2\pi kr) f(r) r dr,$$

where $J_0(kr)$ is the zeroth-order Bessel J-function and $k \in [0, \infty)$. The 2-D Hankel transform can be applied to the OZ equation to obtain

$$\mathcal{H}[h](k, z_1, z_2) = \mathcal{H}[c](k, z_1, z_2) + \int \rho(z_3)\mathcal{H}[h](k, z_1, z_2)\mathcal{H}[c](k, z_1, z_3) dz_3.$$

In the discrete analog, we discretise z, r, and k by setting $z_m = m\Delta z$, $m = 1, \ldots, M$, $r_\ell = \ell \Delta r$, $\ell \in 1, \ldots, K$ and $k_\ell = j_\ell^{(0)} R / j_{K+1}^{(0)}$, where $j_\ell^{(0)}$ is the ℓth root of the zeroth-order Bessel J function, and R is the maximum value at which the estimator, $\hat{h}$ from equation (3.3) is computed. We set $R = 3\mu$m since beyond this distance, the radially symmetric PCF showed little variation. We found that setting the discretisation parameters, $M = 100$ and $K = 256$ yielded a sufficiently fine discretisation. A discrete analogue of the Hankel transform is numerically computed in Matlab using the algorithm devised in [23] for each pair of z_m and z_n with $m, n \in [1, M]$.

Approximating the z-integral with the trapezoid method, the discrete approximation, C_{k_ℓ, z_m, z_n} of $\mathcal{H}[c](k_\ell, z_m, z_n)$ for each fixed value of k_ℓ and z_m, is found by solving

$$H_{k_\ell, z_m, z_n} = C_{k_\ell, z_m, z_n} + \sum_{n'=1}^{M} \hat{\rho}(n'\Delta z) H_{k_\ell, z_n, z_{n'}} C_{k_\ell, z_m, z_{n'}} w_{n'} \Delta z. \tag{4.1}$$

The symbol $w_{n'}$ are the trapezoid rule quadrature weights (i.e., $w_n = 1$ for $n = 2, \ldots, M-1$

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

and $w_1 = w_M = 1/2$). Defining $\mathcal{I}$ as the $M \times M$ identity matrix, and the matrix $\mathcal{M}$ by

$$[\mathcal{M}_k]_{i,j} \equiv \hat{\rho}(i\Delta z)H_{k,z_i,z_j}\Delta z, \quad i, j \in [1 : M].$$

Equation (4.1) can be written as a set of matrix equations

$$\boldsymbol{H}_{k_\ell,z_m,z_n} = (\mathcal{I} + \mathcal{M}_{k_\ell})\boldsymbol{C}_{k_\ell,z_m,z_n} \ \ \forall k_\ell, \forall z_m. \tag{4.2}$$

These matrix equations were solved using LU-decomposition implemented in Matlab. In principle, the matrix $\mathcal{I} + \mathcal{M}$ could be ill-conditioned. However, in the simulations we conducted, the condition number of $\mathcal{I} + \boldsymbol{M}$ ranges only from 1 to 100. We also note that the trapezoid rule approximation is second-order accurate, although this does not imply second-order accuracy of the solution since it does not account for errors in the estimation of h by $\hat{h}$[8].

Finally, after solving (4.2) for each value of ℓ and m, the Hankel transform can be applied to obtain c_{r_ℓ,z_m,z_n}, the discrete approximation of $c(r_\ell, z_m, z_n)$. If there are K radial nodes, and M vertical nodes, KM equations must be solved, and $2M^2$ transforms and inverse transforms must be computed. Although this leads to poor scaling, since we only need to compute the DCF once, and can then store its value, the cost is not prohibitive, and, for the values of K and M we use, can be found in under a minute. In fact, the computation of $\hat{g}(r)$ is, by a substantial amount, the most time consuming step, followed by the discrete Hankel transformations of h and c for each z_m, z_n pair. The resulting pair potential for data set #3 is shown in Figure 6.

After computing c, a discrete approximation of the singlet energy now can be obtained. Discretising equation (2.5) with the trapezoid rule, and using the estimate, c_{r_ℓ,z_m,z_n}, we arrive at a forward Euler approximation of $\phi(z)$

$$\frac{\left(\phi_{z_n} + \log\hat{\rho}(z_n)\right) - \left(\phi_{z_{n-1}} + \log\hat{\rho}(z_{n-1})\right)}{\Delta z} = 2\pi\left(\sum_{m=1}^{M}\left(\sum_{\ell=1}^{K} c_{r_\ell,z_m,z_n}\Delta r_\ell\right)\hat{\rho}_z(m\Delta z)w_m\Delta z\right),$$

with trapezoid rule weights, w_m defined as, $w_m = 1$ for $2 \leqslant m \leqslant M-1$ and $w_m = 0.5$ for $m = 1$ and $m = M$. It turns out that the initial value, $\phi(z_0)$ is arbitrary, so is set to 0.

Given $\hat{\rho}'(z)$, the discretisation should be second-order accurate in Δz and Δr_ℓ. As with the estimation of $c(|\boldsymbol{r}_1 - \boldsymbol{r}_2|_{xy}, z_1, z_2)$, a rigorous error derivation must take into account error terms due to approximations used in estimating $\rho(z)$ by $\hat{\rho}(z)$ and $c(|\boldsymbol{r}_1 - \boldsymbol{r}_2|_{xy}, z_1, z_2)$ by c_{r_ℓ,z_m,z_n}. For the approximation of $\phi(z)$, there is no advantage to using Hankel transforms for the radial integral since the real-space computation is explicit in $\phi(z)$. This differs from the pair energy computation where the real-space integral equation is a 3-D integral equation but, the transformed equation is a set of decoupled 1-D integral equations. In that case, transforming in the r-coordinate is highly beneficial. We also note that other numerical quadrature methods could be applied to compute the values of the integrals; however, we expect that the dominant error source is from the approximation of the

[8] The matrix norm used here is the 2-norm. This norm is also used in the computation of the condition number.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

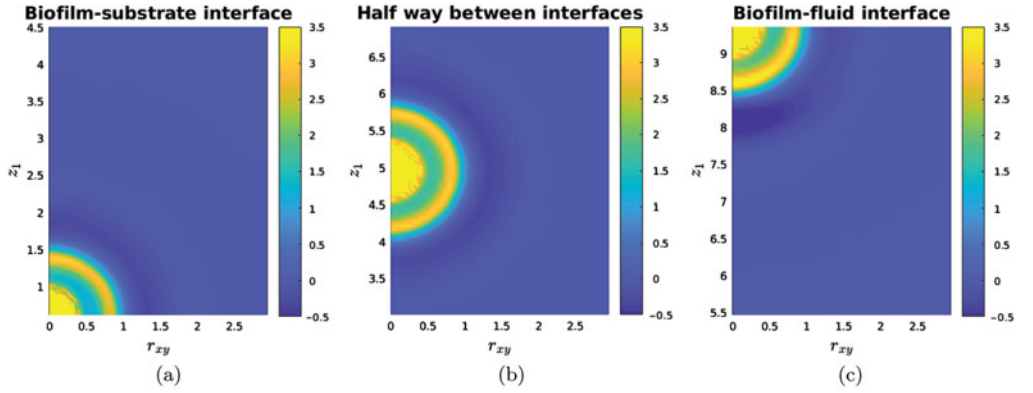

FIGURE 6. The potential, $v(r_{xy}, z_1, z_2)$ as it is computed from the hypernetted-chain equation. The three subfigures show $v(r_{xy}, z_1, z_2)$ for three different values of z_2. Although the pair potential at all values is similar, they are not equivalent; the peaks and valleys in the pair correlation in (a) are slightly more pronounced than in (b) and (c). The differences in the pair potential at different heights influences the height-dependent number density trends we observed in the biofilm data.

PCF and spatially variable number density. Thus, we do not expect significant differences resulting from different quadrature choices.

5 A Metropolis–Hastings algorithm for generating biofilm realisations

The pair energy function and singlet potential form the basis for a MH algorithm to generate 'artificial' biofilms. The practical aspect of MH (and more generally MCMC) algorithms that makes them so useful is that they rely only on unnormalised probability densities. Given two realisations of the PIM, denoted $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, each containing n points in W, and an unnormalised probability density, $q^{(n)}(\boldsymbol{X}_1) \propto f^{(n)}(\boldsymbol{X}_1)$, the ratio

$$\frac{q^{(n)}(\boldsymbol{X}_1)}{q^{(n)}(\boldsymbol{X}_2)} = \frac{f^{(n)}(\boldsymbol{X}_1)}{f^{(n)}(\boldsymbol{X}_2)} = \frac{\exp\left[-\sum_i \phi(z_i^{(1)}) - \sum_{i<j} v(|\boldsymbol{r}_i^{(1)} - \boldsymbol{r}_j^{(1)}|, z_i, z_j)\right]}{\exp\left[-\sum_i \phi(z_i^{(2)}) - \sum_{i<j} v(|\boldsymbol{r}_i^{(2)} - \boldsymbol{r}_j^{(2)}|, z_i, z_j)\right]} \tag{5.1}$$

can be computed efficiently in comparison to evaluating or approximating the probability density, $f^{(n)}(\boldsymbol{X})$. If $\boldsymbol{X}_2$ differs from $\boldsymbol{X}_1$ by the location of only one point such that $\boldsymbol{X}_1 = \boldsymbol{X}_2 \backslash \{\boldsymbol{r}\} \cup \boldsymbol{r}'$, then the quantity in equation (5.1) simplifies and can be computed in $\mathcal{O}(n)$ operations as

$$\frac{f^{(n)}(\boldsymbol{X}_1)}{f^{(n)}(\boldsymbol{X}_2)} = \exp\left[-(\phi(z') - \phi(z)) - \sum_{i=1}^{n} v(|\boldsymbol{r}_i - \boldsymbol{r}|_{xy}, z_i, z) + \sum_{i=1}^{n} v(|\boldsymbol{r}_i - \boldsymbol{r}'|_{xy}, z_i, z')\right]. \tag{5.2}$$

Equation (5.2) can be used to generate realisations of the PIM conditional on there being n points in the domain. Such realisations would belong to the so-called canonical ensemble. However, in practice, it is often more favourable to generate realisations that are not conditional on n. Such a method samples from the grand-canonical ensemble. A MH algorithm with both move, and birth–death steps [19, 33] is a practical method of

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

Algorithm 1 MH algorithm for generating realisations of a point process with spatial characteristics similar to those of the experimental data in a domain W, of finite size. As a convergence criterion, we compute the norm of the difference in the PCF of $\boldsymbol{X}_k$ and the experimental data.

Generate n_0 randomly placed points in a window, W. Denote this set of points $\boldsymbol{X}_0$.
Generate a random number $q \sim \text{Uniform}([0,1])$
if $q < 0.5$ **then**
  Displace one point, $\boldsymbol{x}_i \in \boldsymbol{X}_k$, chosen at random, by a uniformly random displacement, $\delta\boldsymbol{x}$ and set

$$\tilde{\boldsymbol{X}} = \left\{ \left(\boldsymbol{X}_k \backslash \{\boldsymbol{x}_i\} \right) \cup \{\boldsymbol{x}_i + \delta\boldsymbol{x}\} \right\}.$$

  Compute

$$\alpha_m = \min\left(\frac{f^{(n)}(\tilde{\boldsymbol{X}})}{f^{(n)}(\boldsymbol{X}_k)}, 1 \right)$$

  according to equation (5.2) using interpolation as needed.
  if $\tilde{\boldsymbol{X}} \subset W$ **then**
    With probability α_m, set $\boldsymbol{X}_{k+1} = \tilde{\boldsymbol{X}}$.
  end if
else
  Generate a random number $p \sim \text{Uniform}([0,1])$
  if $p < 0.8$ **then**
    Generate a point $\boldsymbol{\xi} \sim \text{Uniform}(W_{\boldsymbol{X}_k})$ in region $W_{\boldsymbol{X}_k} = \{\boldsymbol{\xi} | \min |\boldsymbol{\xi} - \boldsymbol{X}_k| > 0.35\}$
    Compute

$$\alpha_b = \min\left(\frac{f^{(n+1)}(\boldsymbol{X}_k \cup \boldsymbol{\xi})}{f^{(n)}(\boldsymbol{X}_k)} \frac{\int \rho^{(1)}(\boldsymbol{r}) d\boldsymbol{r}}{n+1}, 1 \right)$$

    With probability α_b, $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k \cup \boldsymbol{\xi}$
  else
    Choose a number $m \sim \text{Uniform}([1:n])$
    Compute

$$\alpha_d = \min\left(\frac{f^{(n-1)}(\boldsymbol{X}_k \backslash \{\boldsymbol{x}_m\})}{f^{(n)}(\boldsymbol{X}_k)} \frac{n}{\int \rho^{(1)}(\boldsymbol{r}) d\boldsymbol{r}}, 1 \right)$$

    With probability α_d, set $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k \backslash \{\boldsymbol{x}_m\}$.
  end if
end if
Set $k \leftarrow k+1$ and repeat steps 2–4 unless a *convergence criterion* has been reached. If a convergence criterion has been reached, output $\boldsymbol{X}_{k+1}$ and exit.

obtaining such realisations. Given $q^{(n)}$ (or $f^{(n)}$), the MH algorithm proceeds with the steps detailed in Algorithm 1.

Because $\phi(z)$ and $v(|\boldsymbol{r}_1 - \boldsymbol{r}_2|_{xy}, z_1, z_2)$ are computed using the methods of Section 4 on a grid and are not known analytically, interpolation is used to approximate their values at the points, $\boldsymbol{r}_i \in \Phi$, which are arbitrary points in space. Linear interpolation was the chosen interpolation method because it is computationally inexpensive and accurate in our case since the grid of points on which the energy is computed on in the previous section has

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

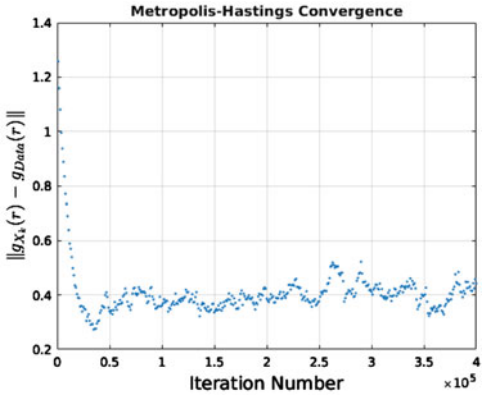

FIGURE 7. Convergence of the Metropolis–Hastings algorithm described in Algorithm 1.The convergence diagnostic used was the norm of the difference in the pair correlation computed from $\boldsymbol{X}_k$ and the data set. This diagnostic was computed every 1000th iteration. Even after the burn-in phase, there is still variability in the energy between iterations, but there is no downward trend diagnostic.

a small spatial step. Potential improvements could apply the ideas in [8] obtain Markov chains such that have reduced dependence on the choice of interpolation method. For boundary conditions, we imposed periodicity in the two horizontal directions, and an impenetrable boundary at the top and bottom of the domain, by setting $\phi(z) = \infty$ for points when a proposed move sets a point outside of W.

The birth-step we employ is somewhat unconventional, but still well defined in the context of MH algorithms. We found that the mean number of points in realisations of the algorithm depend strongly on p. Setting $p = 0.82$ and generating points in the region greater than 0.35 μm away from points in $\boldsymbol{X}$ (due to the hard-sphere property of $v(\cdot,\cdot)$) lead to point processes with mean approximately equal to the number of points in the experimental data set. The factor $(n+1)/\int \rho^{(1)}(\boldsymbol{r})d\boldsymbol{r}$ is the ratio of probabilities of there being $n+1$ to n points in a Poisson point process. For PIM processes, this factor is intractable; thus, we use the easily computable Poisson probability ratio in its place. Note that although we simulate in the grand canonical ensemble, we still assume a fixed volume based on the experimental data.

In step 17 of Algorithm 1, a halting criterion must be used to determine whether to continue finding $\boldsymbol{X}_{k+1}$ or to exit. Such criteria are usually based on easily computable quantities such as the total unnormalised density, $q^{(n)}(\boldsymbol{X})$, or, a characteristic such as the empirical PCF [33, Section 8]. In our case, we observed that, with an update step drawn from a uniform distribution, $\delta\boldsymbol{X} \propto U([-0.5, 0.5]^3)$, the total energy levelled off after $\mathcal{O}(10^5)$ iterations, and the average number of acceptances was approximately $1/2$ of the number of attempts. The computed value of $g(r)$ for each sample also stabilised by this point in each case. Thus, as a halting criterion, we compute the total energy of the sample every 10,000 iterations and stop after it has levelled off. In practice, this lead to convergence within 500,000 iterations in each simulation that we conducted.

The convergence in total energy of one particular realisation is shown in Figure 7. There is a clearly distinguishable initial period of decreasing energy followed by a valley where the energy remains within a small range.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

6 Comparison of material properties

Although experimental results on biofilms often range drastically between different studies, it is generally agreed that over short time scales and moderate mechanical stresses, biofilms behave as viscoelastic materials [36]. One way to characterise biofilm rheology is through measurement of the dynamic moduli discussed below.

Under mild restrictions [6], the stress, σ in a linear viscoelastic material undergoing small-amplitude oscillatory shear strain deformation at frequency, ω and amplitude, ϵ_0 can be written as

$$\sigma(\omega, t) = \epsilon_0 G^*(\omega) e^{i\omega t}.$$

The symbol G^* is known as the complex modulus. The storage and loss moduli (collectively known as the dynamic moduli), $G'(\omega)$ and $G''(\omega)$ are the real and imaginary parts:

$$G^*(\omega) = G'(\omega) + iG''(\omega).$$

Writing $\sigma(\omega) = \sigma_0(\omega) e^{i\omega t - i\delta(\omega)}$, where the $\delta(\omega)$ is the phase lag between the stress and strain, we arrive at

$$G'(\omega) = \frac{\sigma_0(\omega)}{\epsilon_0} \cos(\delta(\omega)) \quad G''(\omega) = \frac{\sigma_0(\omega)}{\epsilon_0} \sin(\delta(\omega)).$$

Given $\boldsymbol{X} \sim PIM(\phi, v)$, the stress and strain in a biofilm undergoing oscillatory shear deformation [36] can be measured as functions of time using the hrIBM discussed in [25,43]. This method yields a finite difference approximation of the velocity field, and an approximation of the forces exerted by a biofilm. The strain can be estimated by considering the motion of tracer particles in the fluid, and the stress is measured by averaging the forces over the top part of the biofilm, and then dividing by the area of the top of the biofilm. The phase lag is approximated by observing when the maxima in the stress and strain occur in time

$$\delta(\omega) = \arg \max_{t \in [0, 2\pi/\omega]} \sigma(\omega, t) - \arg \max_{t \in [0, 2\pi/\omega]} \epsilon(\omega, t).$$

In Figure 8, we depict the dynamic moduli, for four different types of point processes and the experimental data. Since we have four experimental data sets available to us, we repeat the comparison for each data set. The 'random' data sets are realisations of a Poisson process of constant number density, equal to the average number density of the experimental data sets. The 'model' data is based on realisations of the model in Section 4, the grid-aligned data use a regular grid of points in simulation, and the grid-aligned plus perturbation uses the same grid-aligned data with an added perturbation drawn from a normal distribution of mean 0 and variance 0.2 μm added to each point.

We see that the Poisson point process and PIM models yield dynamic moduli in closer agreement to the experimental data than a grid aligned approximation, although some amount of improvement is seen when the grid aligned data is randomly perturbed. In each case, the comparisons are made between point processes with approximately the same number of points as the experimental data. (In the uniformly random, and model based simulations, the number of points is exactly the same as the experimental data.)

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

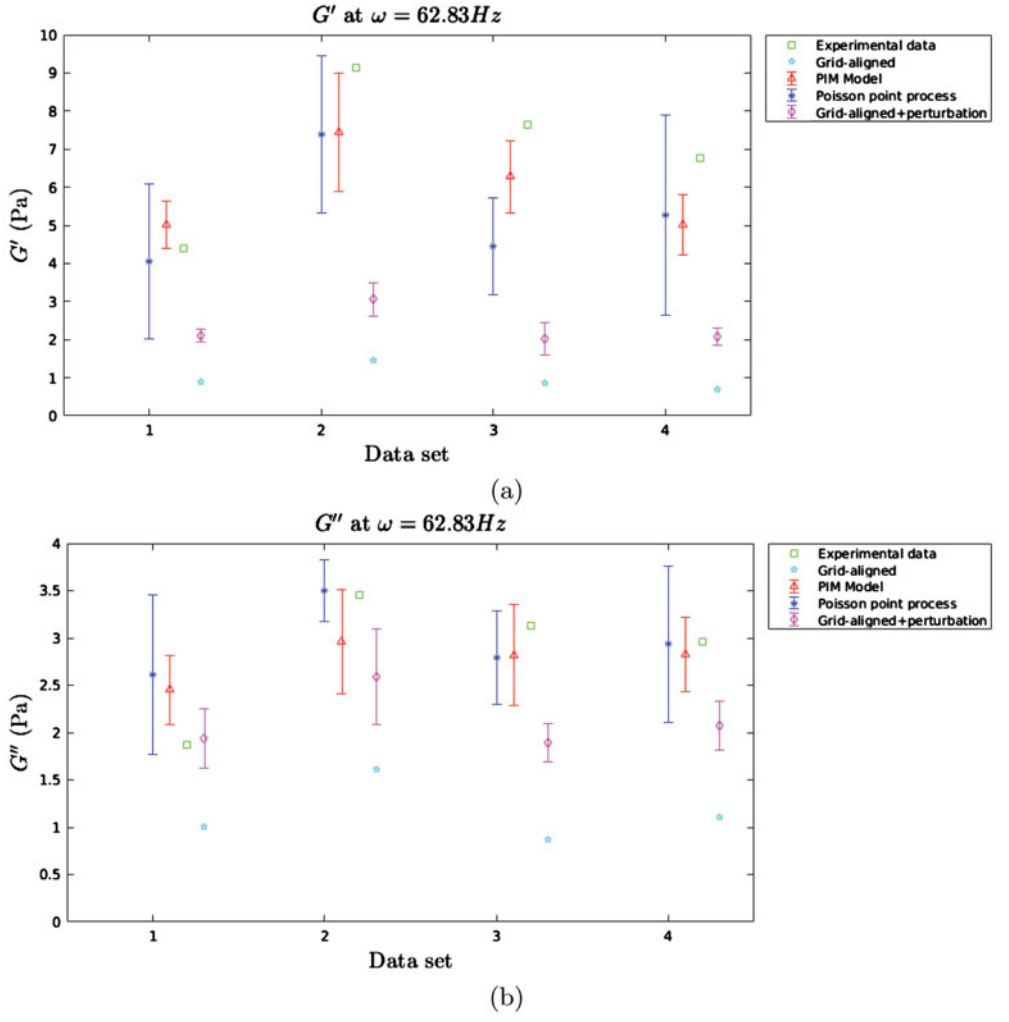

FIGURE 8. Comparison of material properties between experimental data and simulated data sets. Panel (a) shows the comparison between the statistical models and the experimental data for G', and panel (b) shows the the comparison for G''. Both figures are from simulations at $\omega = 62.83$ Hz. For the grid aligned data, the data falls on regular grid of the form $\boldsymbol{X} = (ih, jh, kh)$, for i, j, k being integers. For the Poisson point process data (e.g., random data) and PCF Model, each entry corresponds to the average of five samples containing the same number of points and domain as each of four experimental data sets. For the grid aligned plus perturbation, the results are from five trials with the grid aligned data and a perturbation drawn from a normal distribution of mean 0 and variance of 0.2 μm. The x-axis numbering corresponds to each of the four data sets the statistical models were designed to compare with. Error bars in this figure are 95% confidence intervals of the posterior distribution of dynamic moduli observed from each point process model. These are computed using Matlab's *normfit* function.

To estimate the variability in the dynamic moduli for each of the point processes we have mentioned, the dynamic moduli are computed for five different realisations of each point process. Although it would be beneficial to observe a larger number of trials over a

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

range of frequencies, the computational cost of such an endeavour is currently prohibitive. Each hrIBM simulation on a computer with an Intel i7-7600U CPU with clock rate 2.8 GHz and 16 GB of RAM takes 6–8 hr.

7 Discussion

In [2,25,43,52], systems of points immersed in a viscous fluid and connected by a network of viscoelastic fibres are used as models of complex microbiological fluids. The points in each application are distributed according to different procedures. By placing such point distributions in the context of point process theory, we strive to elucidate the effect that the positions of points have on the rheology of the resulting system. In particular, in this work, we show that for biofilms (i) empirical data yields biofilms that are exhibit larger dynamic moduli than biofilms artificially generated through simple point process models, (ii) the empirically informed PIM leads to realisations with statistical characteristics that strongly match empirical data, and dynamic moduli that closely match data, and (iii) the Poisson point process model leads to close agreement with empirical dynamic moduli, but fails to match statistical characteristics of the empirical data.

In contrast to our results showing variation in the dynamic moduli depending on the point process model, we note that in a intriguing article [52], the network topology of an immersed boundary method model was not observed to have strong effects on bulk viscoelastic properties of the system. We believe that this difference with our findings is due to the magnitude of the number density of the point process being simulated and how it compares to the average connectivity of each point. In our simulations, a bacterium is, on average, connected to approximately 9 other bacteria through viscoelastic linkages, whereas [52] used 2 regular networks of points with 14 and 27 linkages per point, respectively. We conjecture that for high-connectivity networks, the rheological properties of the material are less sensitive to the spatial positioning of points, whereas lower connectivity networks exhibit more sensitivity. In a future work, we plan to further study this idea.

Another future avenue of research would be to integrate a model for the network topology (i.e., given bacteria at $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, what is the chance that there is a viscoelastic linkage connecting them?) into the point process model. The bacteria in a biofilm are typically interconnected through a complicated network of viscoelastic linkages whose properties effect the bulk material properties of the biofilm. Currently, to model this network, we use a deterministic rule specifying that if a pair of bacteria is separated by a distance less than a cutoff radius, r_c they are linked, and those separated by more than r_c are not linked. This network topology was originally proposed in [2], and was again used in [25, 43]. Although simple, it yields realistic results when combined with appropriate viscoelastic models of the linkages [43]. The reason that we do not introduce a statistical model for the network topology is that current state-of-the-art experiments allow for determination of bacteria locations, but to the extent of our knowledge, there is no data on the connectivity of bacteria in a biofilm. We are very interested in analysing connectivity statistics should such data become available.

The results in Section 6 show how different models of the positions of bacteria in a biofilm can influence the mechanical properties of the simulated biofilm. It was found

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

that the Poisson process model, and grid-aligned model do not yield results that are consistent with statistical characteristics (e.g., Figure 4(a) and (b)) of the experimental data. However, we were surprised to find that despite this difference from the experimental data, the Poisson model yields biofilms with similar dynamic moduli, performing as well as the PIM in terms of recreating the mechanical properties of a biofilm. In contrast, the grid-aligned model, as shown in Figure 8 is not a close match. These findings indicate that nonuniformity can lead to stronger biofilms in comparison to the grid-aligned case. We also show that the model introduced in Section 4, with first- and second-order characteristics informed by experimental data, yields agreement in the material properties and agreement in statistical properties of the data.

Although the Poisson point process model exhibits similar dynamic moduli as experimental data, from a physical interpretation, it is clear that the bacteria positions cannot conform to a Poisson process. The lack of correlations between point locations, a defining feature of Poisson processes, implies that for any radius, r, there is a non-zero probability that within a given realisation, two points of the process are separated by a distance less than r. In contrast, bacteria have finite radii and are impenetrable; thus, the centres of mass of two bacteria cannot be separated by less than some hard-sphere radius. The hard-sphere property of the experimental data is readily apparent upon computation of the nearest neighbour distribution and the PCF. However, we note that the presence of an initial low magnitude peak in the PCF is not typical of hard-sphere processes, and requires a more complicated model such as the empirically informed PIM model we have proposed.

Although not the focus of our work, an interesting result that arose in the course of this study was the observation of variations in the number density of bacteria near the fluid-biofilm interface. It would be interesting to discover whether this phenomenon is a passive effect due to fluid motion at the biofilm-fluid surface or other experimental conditions, or whether it is a strategy employed by biofilm forming bacteria to improve the biological fitness of a given biofilm.

Acknowledgements

The authors would also like to thank Mike Solomon (University of Michigan) and John Younger (Akadeum Life Sciences) for insightful discussions and suggestions concerning this work. Furthermore, we would like to thank the anonymous reviewers for many helpful comments and suggestions.

References

[1] Abramson, I. S. (1982) On bandwidth variation in kernel estimates-a square root law. *Ann. Stat.* **10**(4), 1217–1223.

[2] Alpkvist, E. & Klapper, I. (2008) Description of mechanical response including detachment using a novel particle model of biofilm/flow interaction. *Water Sci. Technol.* **55**(8–9), 265–273.

[3] Baddeley, A. & Turner, R. (2000) Practical maximum pseudolikelihood for spatial point patterns. *Aust. N. Z. J. Stat.* **42**(3), 283–322.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

[4] Baddeley, A. J., Møller, J. & Waagepetersen, R. (2000) Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Stat. Neerlandica* **54**(3), 329–350.

[5] Billingsley, P. (2008) *Probability and Measure*, John Wiley & Sons, New York.

[6] Christensen, R. M. (1982) *Theory of Viscoelasticity: An Introduction*, 2nd ed, New York: Academic Press.

[7] Coeurjolly, J. F., Møller, J. & Waagepetersen, R. (2017) A tutorial on Palm distribution for spatial point processes. *Int. Stat. Rev.* **85**(3), 404–420.

[8] Conrad, P. R., Marzouk, Y. M., Pillai, N. S. & Smith, A. (2016) Accelerating asymptotically exact MCMC for computationally intensive models via local approximations. *J. Am. Stat. Assoc.* **111**(516), 1591–1607.

[9] Courant, R. & Hilbert, D. (1954) Methods of mathematical physics, Vol. I. *Phys. Today* **7**(5), 17–17.

[10] Crocker, J. C. & Grier, D. G. (1996) Methods of digital video microscopy for colloidal studies. *J. Colloid Interface Sci.* **179**(1), 298–310.

[11] Cronie, O. & van Lieshout, M. N. M. (2018) A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*.

[12] Daley, D. J. & Vere-Jones, D. (2007) *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*, Springer Science & Business Media, New York.

[13] Dzul, S. P., Thornton, M. M., Hohne, D. N., Stewart, E. J., Shah, A. A., Bortz, D. M., Solomon, M. J. & Younger, J. G. (2011) Contribution of the Klebsiella pneumoniae capsule to bacterial aggregate and biofilm microstructures. *Appl. Environ. Microbiol.* **77**(5), 1777–1782.

[14] Epanechnikov, V. A. (1969) Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14**(1), 153–158.

[15] Fai, T. G., Leo-Macias, A., Stokes, D. L. & Peskin, C. S. (2017) Image-based model of the spectrin cytoskeleton for red blood cell simulation. *PLoS Comput. Biol.* **13**(10), e1005790.

[16] Flemming, H. C. (2011) Microbial biofouling: Unsolved problems, insufficient approaches, and possible solutions. In: H.-C. Flemming, J. Wingender, U. Szewzyk (editors), *Biofilm Highlights*, Springer, pp. 81–109.

[17] Gaboriaud, F., Gee, M. L., Strugnell, R. & Duval, J. F. L. (2008) Coupled electrostatic, hydrodynamic, and mechanical properties of bacterial interfaces in aqueous media. *Langmuir* **24**(19), 10988–10995.

[18] Gangopadhyay, A. & Cheung, K. (2002) Bayesian approach to the choice of smoothing parameter in kernel density estimation. *J. Nonparametric Stat.* **14**(6), 655–664.

[19] Geyer, C. J. & Møller, J. (1994) Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Stat.* **21**(4), 359–373.

[20] Guan, Y. (2007) A least-squares cross-validation bandwidth selection approach in pair correlation function estimations. *Stat. Probab. Lett.* **77**(18), 1722–1729.

[21] Guan, Y. (2008) On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *J. Am. Stat. Assoc.* **103**(483), 1238–1247.

[22] Guélon, T., Mathias, J. D. & Stoodley, P. (2011) Advances in biofilm mechanics. In: H.-C. Flemming, J. Wingender, U. Szewzyk (editors), *Biofilm Highlights*, Springer, pp. 111–139.

[23] Guizar-Sicairos, M. & Gutiérrez-Vega, J. C. (2004) Computation of quasi-discrete Hankel transforms of integer order for propagating optical wave fields. *J. Opt. Soc. Am. A* **21**(1), 53–58.

[24] Hall, P. & Marron, J. S. (1991) Local minima in cross-validation functions. *J. R. Stat. Soc. Ser. B (Methodological)* **53**(1), 245–252.

[25] Hammond, J. F., Stewart, E. J., Younger, J. G., Solomon, M. J. & Bortz, D. M. (2014) Variable viscosity and density biofilm simulations using an immersed boundary method, Part I: Numerical scheme and convergence results. *Comput. Model. Eng. Sci.* **98**(3), 295–340.

[26] Hansen, J. P. & McDonald, I. R. (1990) *Theory of Simple Liquids*, Elsevier, London, UK.

[27] Hardle, W., Marron, J. S. & Wand, M. P. (1990) Bandwidth choice for density derivatives. *J. R. Stat. Ser. B (Methodological)* **52**(1), 223–232.

[28] Jones, M. C. (1993) Simple boundary correction for kernel density estimation. *Stat. Comput.* **3**(3), 135–146.

[29] Kerscher, M., Szapudi, I. & Szalay, A. S. (2000) A comparison of estimators for the two-point correlation function. *Astrophys. J. Lett.* **535**(1), L13.

[30] Landy, S. D. & Szalay, A. S. (1993) Bias and variance of angular correlation functions. *Astrophys. J.* **412**, 64–71.

[31] Laspidou, C. S. & Rittmann, B. E. (2004) Modeling the development of biofilm density including active bacteria, inert biomass, and extracellular polymeric substances. *Water Res.* **38**(14), 3349–3361.

[32] Lovett, R., Mou, C. Y. & Buff, F. P. (1976) The structure of the liquid-vapor interface. *J. Chem. Phys.* **65**, 2377.

[33] Moller, J. & Waagepetersen, R. P. (2003) *Statistical Inference and Simulation for Spatial Point Processes*, CRC Press, Boca Raton, FL.

[34] Ornstein, L. S. & Zernike, F. (1914) The influence of accidental deviations of density on the equation of state. *Koninklijke Nederlandsche Akademie van Wetenschappen Proceedings* **19**(2), 1312–1315.

[35] Parzen, E. (1962) On estimation of a probability density function and mode. *Ann. Math. Stat.* **33**(3), 1065–1076.

[36] Pavlovsky, L., Younger, J. G. & Solomon, M. J. (2013) In situ rheology of Staphylococcus epidermidis bacterial biofilms. *Soft Matter* **9**(1), 122–131.

[37] Ripley, B. D. (1991) *Statistical Inference for Spatial Processes*, Cambridge University Press.

[38] Rosenblatt, M. *et al.* (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **27**(3), 832–837.

[39] Silverman, B. W. (1981) Using Kernel density estimates to investigate multimodality. *J. R. Stat. Soc.* **43**(1), 97–99.

[40] Sobczyk, K. & Kirkner, D. J. (2012) *Stochastic Modeling of Microstructures*, Springer Science & Business Media, Boston, MA.

[41] Stewart, E. J., Ganesan, M., Younger, J. G. & Solomon, M. J. (2015) Artificial biofilms establish the role of matrix interactions in Staphylococcal biofilm assembly and disassembly. *Sci. Rep.* **5**, 13081; doi: 10.1038/srep13081.

[42] Stewart, E. J., Satorius, A. E., Younger, J. G. & Solomon, M. J. (2013) Role of environmental and antibiotic stress on Staphylococcus epidermidis biofilm microstructure. *Langmuir* **29**(23), 7017–7024.

[43] Stotsky, J. A., Hammond, J. F., Pavlovsky, L., Stewart, E. J., Younger, J. G., Solomon, M. J. & Bortz, D. M. (2016) Variable viscosity and density biofilm simulations using an immersed boundary method, Part II: Experimental validation and the heterogeneous rheology-IBM. *J. Comput. Phys.* **317**, 204–222.

[44] Stoyan, D., Bertram, U. & Wendrock, H. (1993) Estimation variances for estimators of product densities and pair correlation functions of planar point processes. *Ann. Inst. Stat. Math.* **45**(2), 211–221.

[45] Stoyan, D., Kendall, W. S. & Mecke, J. (1995) *Stochastic Geometry and its Applications*, Akademie-Verlag, Berlin.

[46] Sudarsan, R., Ghosh, S., Stockie, J. M. & Eberl, H. J. (2016) Simulating biofilm deformation and detachment with the immersed boundary method. *Commun. Comput. Phys.* **19**(3), 682–732.

[47] Szapudi, I. & Szalay, A. S. (1998) A new class of estimators for the n-point correlations. *Astrophys. J. Lett.* **494**(1), L41.

[48] Torquato, S. (2013) *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*, Vol. 16, Springer Science & Business Media, New York.

[49] Truskett, T. M., Torquato, S. & Debenedetti, P. G. (1998) Density fluctuations in many-body systems. *Phys. Rev. E* **58**(6), 7369.
[50] Vo, G. D., Brindle, E. & Heys, J. (2010) An experimentally validated immersed boundary model of fluid–biofilm interaction. *Water Sci. Technol.* **61**(12), 3033–3040.
[51] Wand, M. P. & Jones, M. C. (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Am. Stat. Assoc.* **88**(422), 520–528.
[52] Wróbel, J. K., Cortez, R. & Fauci, L. (2014) Modeling viscoelastic networks in stokes flow. *Phys. Fluids (1994-present)* **26**(11), 113102.
[53] Yeong, C. L. Y. & Torquato, S. (1998) Reconstructing random media. *Phys. Rev. E* **57**(1), 495.
[54] Zhang, T., Cogan, N. G. & Wang, Q. (2008) Phase field models for biofilms. I. Theory and one-dimensional simulations. *SIAM J. Appl. Math.* **69**(3), 641–669.
[55] Zhang, T., Cogan, N. G. & Wang, Q. (2008) Phase field models for biofilms. ii. 2-d numerical simulations of biofilm-flow interaction. *Commun. Comput. Phys* **4**(1), 72–101.
[56] Zhao, J., Shen, Y., Haapasalo, M., Wang, Z. & Wang, Q. (2016) A 3d numerical study of antimicrobial persistence in heterogeneous multi-species biofilms. *J. Theor. Biol.* **392**, 83–98.

Appendix

A.1 Bandwidth selection for the estimation of the number density

In Section 4, it was necessary to estimate the number density of biofilms samples. In order complete this task, we used a kernel density based estimator of the form

$$\hat{\rho}(z;b) = \frac{1}{A} \sum_{\boldsymbol{r}_i \in \Phi \cap W} \frac{k_b(z - \hat{\boldsymbol{e}}_z \cdot \boldsymbol{r}_i)}{c(z;b)}.$$

As is typical with kernel density estimation, a choice of the bandwidth, b, must be made. Several typical techniques are discussed in [51]. The general idea is to minimise the *mean integrated square error* (MISE),

$$MISE(\hat{\rho}(z;b)) = \int (\hat{\rho}(z;b) - \rho(z))^2 \, dz.$$

The difficulty is of course the lack of knowledge of $\rho(z)$. Although the choice of b is partly intuitive, (e.g., too large a value leads to an overly smooth estimate, and too small a value leads to an overly jagged estimate), it is difficult to judge the best value among reasonable values of b by mere qualitative observation. Although there are numerous methods of bandwidth selection, we choose to use the LSCV method described in [21] as a first estimate. We also found that by visual examination values of b in the range (0.13, 0.3) seem to provide reasonable results, and thus expect any optimisation method to yield a value in this range. From [21], we optimise

$$LSCV(b) = \int_W \hat{\rho}(z;b)^2 dz - 2 \sum_{\boldsymbol{r}_i \in \Phi \cap W} \hat{\rho}(z_i;b) - k(0;b)/(Ac(z_i;b))$$

$$= \frac{1}{A^2} \sum_{\boldsymbol{r}_i \in \Phi \cap W} \sum_{\boldsymbol{r}_j \in \Phi \cap W} \int_W \frac{k_h(z - z_i) k_h(z - z_j)}{c_b(z)^2} dz - \frac{2}{A} \sum_{\boldsymbol{r}_i \in \Phi \cap W} \sum_{\boldsymbol{r}_j \in \Phi \cap W \setminus \boldsymbol{r}_i} \frac{k_b(z_i - z_j)}{c_b(z_i)}.$$

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

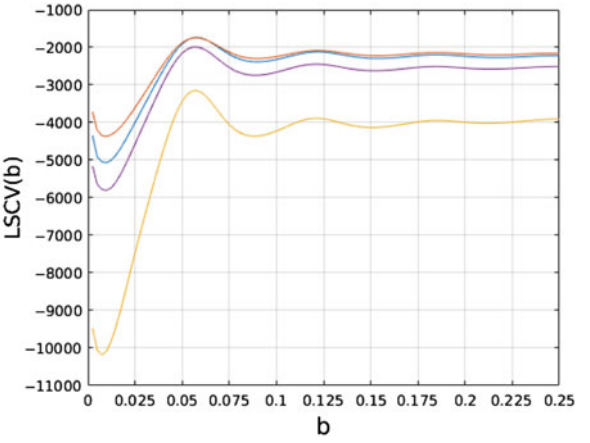

FIGURE A.1. The least squares cross validation value for selected values of b over four data sets.

With the Epanechnikov kernel, and ignoring edge effects for simplicity

$$I(z_i;b) \equiv \int_W k_b(z-z_i)k_b(z)dz = \begin{cases} \frac{3}{160b}\left(32 - 40(z_i/b)^2 + 20|z_i/b|^3 - |z_i/b|^5\right) & |z_i| \leqslant 2b \\ 0 & |z_i| > 2b \end{cases}$$

$$LSCV(b) = \frac{1}{A}\left(\frac{3}{5b}\Phi(W) + \sum_{\boldsymbol{r}_i \neq \boldsymbol{r}_j} I(z_i - z_j;b)\right) - \frac{2}{A}\sum_{\boldsymbol{r}_i \neq \boldsymbol{r}_j} k_b(z_i - z_j).$$

A plot of $LSCV(b)$ versus b is shown in Figure A.1. It can be seen that there exists several minima in each case, and the question is how to choose the 'best' minima. In each case, the first minima, which is also the global minimum, is clearly too small leading to density estimates with unacceptably high variance. We found that the shallow local minimum at $b \approx 0.21$ worked best in practice. We additionally implemented the log-likelihood estimator described in [11] and found very similar results. It is unclear if there exists a method that possess a unique minimum in our case. There is also some evidence that spurious local minimisers of LSCV functionals tend to occur at smaller values than the optimal b; thus, choosing the largest local minimiser seems a suitable strategy [24].

To augment the LSCV estimation, we use the fact that over a finite domain, the number density is a multiple of a probability density function, and adapt the Bayesian estimation technique discussed in [18]. To estimate the bandwidth, they choose a prior distribution $\pi(b)$, and compute the posterior distribution,

$$\pi(b|\{z_1,\ldots,z_n\},z) = \frac{\hat{\rho}_b(z)\pi(b)}{\int_{b=0}^{\infty}\hat{\rho}_{b'}(z)\pi(b')db'}.$$

This estimator has the property of being optimised for each particular value of z over which we evaluate $\hat{\rho}_b(z)$. To determine b,

$$\mathbb{E}\left[b|\{z_1,\ldots,z_n\},z\right] = \int_0^{\infty} b\pi\left(b|\{z_1,\ldots,z_n\},z\right)db,$$

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

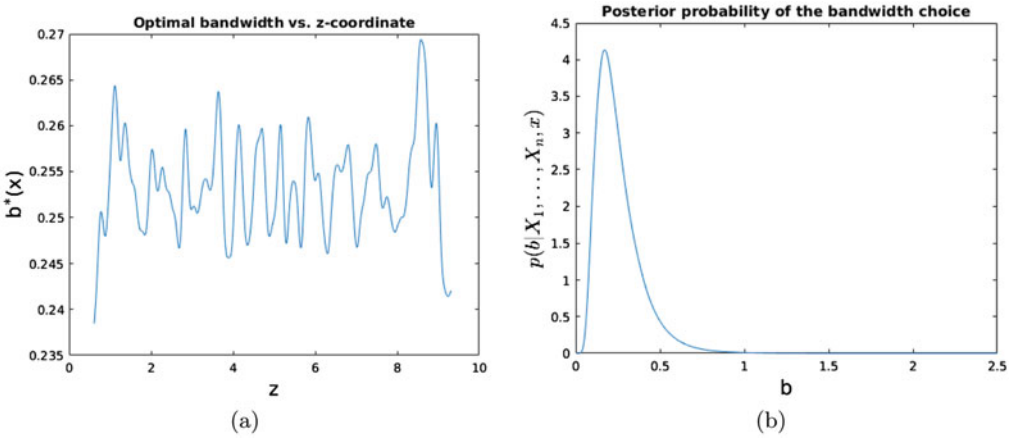

FIGURE A.1. Results of the bandwidth estimation technique [18] applied to number density estimation. On the left, the variability of $b(z)$ over the domain. On the right, the posterior distribution for b for a particular value of z.

is computed. An important ambiguity in this method is the choice of prior, $\pi(b)$. We choose to use a log-normal distribution of the form

$$\pi(b) = \frac{1}{b\sigma\sqrt{2\pi}} e^{-(\log b - \mu)^2/(2\sigma^2)},$$

with $\sigma = 0.5$, and $\mu = -1.5$. Although these choices are clearly somewhat arbitrary, they yield a distribution $\pi(h)$ that has mass where we expect b to be located and tends towards zero rapidly away from that region.

We found that, although the Bayesian estimate seems to do a good job of adapting b for various choices of x, it has a strong tendency to oversmooth if some sort of informed prior estimate for b is absent.

A.2 Bandwidth selection for estimation of the pair correlation function

Similar to the estimation of a number density, the estimation of a PCF bandwidth has an important effect upon the resulting estimator. From a theoretical perspective, it is easiest to analyse estimators of the form

$$\hat{g}(r) = \frac{1}{4\pi r^2 \bar{\gamma}_W(r)\rho^2} \sum_{\boldsymbol{r}_i \in \Phi \cap W} \sum_{\boldsymbol{r}_j \in \Phi \cap W \setminus \{\boldsymbol{r}_i\}} k_b(r - |\boldsymbol{r}_i - \boldsymbol{r}_j|),$$

where $\bar{\gamma}_W(r)$ is known as the *isotropised set covariance* [44,45]. It can be computed as the following integral over $\boldsymbol{t} \in \{\boldsymbol{s} \in \mathbb{R}^3 | \ |\boldsymbol{s}| = r\}$

$$\bar{\gamma}_W(r) = \frac{1}{4\pi r^2} \int v(W \cap W_{\boldsymbol{t}}) d\boldsymbol{t}.$$

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

Although generally intractable to integrate analytically, for boxes, the integral can be computed

$$\begin{aligned}\bar{\gamma}_W(r) =& \frac{1}{4\pi}\int\int v(\mathrm{W}\cap W_{\boldsymbol{t}=(r,\theta,\phi)})\sin\theta \mathrm{d}\phi\mathrm{d}\theta \\ =& \frac{2}{\pi}\int_0^{\pi/2}\int_0^{\pi/2}(l - r\sin\phi\cos\theta)(w - r\sin\phi\sin\theta)(h - r\cos\phi)\sin\theta\, d\theta d\phi \\ =& hlw - \frac{1}{2}(hl + lw + wh)r + \frac{2}{3\pi}(h + l + w)r^2 - \frac{1}{4\pi}r^3,\end{aligned}$$

for $r < \min(h, l, w)$. Then, following [37] and [44]

$$\mathrm{Var}(\hat{g}(r)) \approx \frac{0.6g(r)}{4\pi b\rho^2 r^2\bar{\gamma}_W(r)}.$$

It is noted in [44, Section 5] that this approximation is particularly accurate for hard-core processes. The bias can be computed as

$$\mathrm{Bias}(\hat{g}(r)) = \left(\int k_b(r - r')g(r')dr' - g(r)\right)^2.$$

Since the PCF may exhibit a jump discontinuity near the hard-sphere radius, it is helpful to examine the integrated bias

$$\int \mathrm{Bias}(\hat{g}(r))dr = \int\left(\int k_b(r - r')g(r')dr' - g(r)\right)^2 dr,$$

where $g(r)$ is twice differentiable, the bias and variance can be combined to obtain an approximation for the mean square error as a function of r and b

$$\mathbb{E}\left[(g(r) - \hat{g}(r;b))^2\right] = \frac{0.6g(r)}{4\pi b\rho^2 r^2\bar{\gamma}_W(r)} + \int k_b(r - r')r'^2 dr' g''(r).$$

Assuming $r > b$ since the hard-sphere diameter of bacteria is fairly large in comparison to the range over which we compute $g(r)$

$$\mathbb{E}\left[(g(r) - \hat{g}(r;b))^2\right] = \frac{0.6g(r)}{4\pi b\rho^2 r^2\bar{\gamma}_W(r)} + \frac{1}{5}(b^2 + r^2)g''(r). \tag{A1}$$

Equation (A1) can be minimised for each value of r to yield an analogous result to those typically derived in the case of probability density estimators (cf. [35, 38]),

$$b(r) = \left(\frac{3}{8}\frac{g(r)}{\rho^2 r^2\bar{\gamma}_W(r)g''(r)}\right)^{1/3}.$$

As is the case with density estimation, this expression is of limited usefulness since it depends on the unknown quantities, $g(r)$ and $g''(r)$. However, bandwidth selection methods, such as LSCV, and *biased cross validation* (BCV) [21, 51] can be applied to the integrated MSE to estimate an optimal value of b across the entire interval. For instance, the LSCV estimator for $g(r)$ can be formulated as in [20, equation 4]. In addition, to

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

the minimisation method introduced in [20], we also employ a 'binning' technique to estimate optimal values of b over disjoint portions of the overall interval of computation. The motivation for this adaptation is that the behaviour of $g(r)$ is quite different near the hard-sphere radius in comparison to the asymptotic behaviour for $g(r)$ as r grows. It seems sensible that different bandwidths should be applied in these different regimes. Thus, we employ LSCV functionals of the form

$$LSCV(h;[r_0,r_1]) = 4\pi \int_{r_0}^{r_1} \hat{g}(r;b)^2 r^2 dr - 2 \sum_{r_0 \leqslant |\boldsymbol{r}_i - \boldsymbol{r}_j| \leqslant r_1}^{\neq} \frac{\hat{g}^{-(\boldsymbol{r}_i,\boldsymbol{r}_j)}(|\boldsymbol{r}_i - \boldsymbol{r}_j|;h)}{\bar{\gamma}_W(|\boldsymbol{r}_i - \boldsymbol{r}_j|)\rho(\boldsymbol{r}_i)\rho(\boldsymbol{r}_j)}. \tag{A2}$$

The symbol $\hat{g}^{-(\boldsymbol{r}_1,\boldsymbol{r}_j)}(r;h)$ indicates the computation of the PCF with points $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ ignored. The expectation of the summation in equation (A2) is shown in [20] to converge in the limit of a large domain to

$$\int \hat{g}(r;h) g(r;h) 4\pi r^2 dr.$$

One can see from this the similarity to classical LSCV estimation of bandwidth for kernel density estimation [51].

In practice, we found that the summation of $\hat{g}^{-(\boldsymbol{r}_1,\boldsymbol{r}_2)}(\boldsymbol{r}_1,\boldsymbol{r}_2)$ was too expensive, leading to extremely lengthy computations. To alleviate this issue, we approximated $\hat{g}^{-(\boldsymbol{r}_1,\boldsymbol{r}_2)}(\boldsymbol{r}_1,\boldsymbol{r}_2)$ as

$$\hat{g}^{-(\boldsymbol{r}_1,\boldsymbol{r}_2)}(|\boldsymbol{r}_{12}|) \approx \mathcal{I}\hat{g}(|\boldsymbol{r}_{12}|) - \frac{2}{4\pi|\boldsymbol{r}_{12}|^2\bar{\gamma}(|\boldsymbol{r}_{12}|)} \sum_{\boldsymbol{r}_i \neq \{\boldsymbol{r}_1,\boldsymbol{r}_2\}} \left(k_b(|\boldsymbol{r}_{12}| - |\boldsymbol{r}_{1i}|) + k_b(|\boldsymbol{r}_{12}| - |\boldsymbol{r}_{2i}|)\right),$$

where $\mathcal{I}\hat{g}(\cdot)$ is the linear interpolant of $\hat{g}(\cdot)$ to some value (i.e., $|\boldsymbol{r}_{12}|$). The factor of 2 in the numerator of the second term is due arises since the terms involving $\boldsymbol{r}_{1i}$,$\boldsymbol{r}_{2i}$, $\boldsymbol{r}_{i1}$, and $\boldsymbol{r}_{i2}$ must be subtracted from $\hat{g}(r)$. However, since $|\boldsymbol{r}_{ji}| = |\boldsymbol{r}_{ij}|$, there are only $\Phi(W)$ unique terms in the sum. With this fix, once $\hat{g}(r)$ is known, $\hat{g}^{-(\boldsymbol{r}_1,\boldsymbol{r}_2)}(|\boldsymbol{r}_{12}|)$ can be approximated in $\mathcal{O}(\Phi(W))$ computations as opposed to $\mathcal{O}(\Phi(W)^2)$ computations. As long as $\hat{g}(\cdot)$ is initially computed on a sufficiently dense set of points, the interpolant will be quite accurate. The resulting optimal values of b for different ranges of r are shown in Table A.1. In practice, the value of $b(r)$ is assumed to be piecewise linear in r taking on the reported value in A.1 at the midpoint of each interval. This prevents artificial discontinuous at the end points of the intervals over which b was estimated. Of course, if the estimator is accurate, one would expect such discontinuities to be small. Indeed, when $b(r)$ is a piecewise constant, discontinuities are difficult to discern by sight. We also use numerical integration to compute the relevant integrals in equation (A2).

One final issue with LSCV bandwidth selection is the presence of multiple minima. For the PCF LSCV functional, spurious minima near $b = 0$ were observed. For the probability density estimation, it has been suggested that spurious minima are usually less than the ideal bandwidth [24]. Thus, when multiple minima are present, we choose the minima that is largest over the range of values we consider for b.

An alternative approach to variable bandwidth estimation is use techniques such as that introduced in [1]. We believe such an approach would be effective as well.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press

Table A.1. *Values of b determined through LSCV optimisation are reported for intervals of r*

b	$r_{\min}$	$r_{\max}$
0.050	0.3	0.7
0.3700	0.7	1.6
0.2150	1.6	2.5
0.3960	2.5	5.0

A.3 Inhomogeneity of the direct pair correlation function in non-stationary processes

The motivation for using a transversely anisotropic pair correlation and DCFs was attributed to properties of the Ornstein–Zernike equation. In this section, we demonstrate why the Ornstein–Zernike equation implies a loss of translation invariance in the PCF and DCF when the density is variable.

Consider the inhomogeneous Ornstein–Zernike equation as shown in equation (2.8). Let the two pairs of points $\{\boldsymbol{r}_1, \boldsymbol{r}_2\}$ and $\{\boldsymbol{r}'_1, \boldsymbol{r}'_2\}$ satisfy $\boldsymbol{r}_1 - \boldsymbol{r}_2 = \boldsymbol{r}'_1 - \boldsymbol{r}'_2 = \delta\boldsymbol{r}$, and assume that the pair correlation and DCFs are translation invariant. Then, using the transformation $\boldsymbol{r}'_1 = \boldsymbol{r}_1 + \boldsymbol{x}$, $\boldsymbol{r}'_2 = \boldsymbol{r}_2 + \boldsymbol{x}$

$$h(\boldsymbol{r}_1 - \boldsymbol{r}_2) = c(\boldsymbol{r}_1 - \boldsymbol{r}_2) + \int \rho(\boldsymbol{r}_3) c(\boldsymbol{r}_1 - \boldsymbol{r}_3) h(\boldsymbol{r}_2 - \boldsymbol{r}_3) d\boldsymbol{r}_3$$

$$- \; h(\boldsymbol{r}'_1 - \boldsymbol{r}'_2) = c(\boldsymbol{r}'_1 - \boldsymbol{r}'_2) + \int \rho(\boldsymbol{r}_3) c(\boldsymbol{r}'_1 - \boldsymbol{r}_3) h(\boldsymbol{r}'_2 - \boldsymbol{r}_3) d\boldsymbol{r}_3$$

$$0 = \int \rho(\boldsymbol{r}_3) c(\boldsymbol{r}_1 - \boldsymbol{r}_3) h(\boldsymbol{r}_2 - \boldsymbol{r}_3) d\boldsymbol{r}_3 - \int \rho(\boldsymbol{r}_3) c(\boldsymbol{r}'_1 - \boldsymbol{r}_3) h(\boldsymbol{r}'_2 - \boldsymbol{r}_3) d\boldsymbol{r}_3.$$

In order for this to be true for all translations, $\boldsymbol{x}$, it must be the case that $\rho(\boldsymbol{r}_3) = const$ almost everywhere. Thus, for a smoothly varying number density, it cannot be the case that $\boldsymbol{c}(\boldsymbol{r}_1, \boldsymbol{r}_2)$ and $h(\boldsymbol{r}_1, \boldsymbol{r}_2)$ are both simultaneously translation invariant.

https://doi.org/10.1017/S0956792518000220 Published online by Cambridge University Press