Supplementary information 1 (box) | Primer of some modelling approaches used in microbial ecology

**General characteristics of mathematical models**

The purpose of a model is to simplify reality. Since “a mathematical model is a logical machine for converting assumptions into conclusions”\(^1\), it enforces complete and unambiguous specification of assumptions, which is essential for rigorous testing of hypotheses. An example of a typical simplifying assumption in IBMs is that cells divide instantaneously when they reach a threshold volume; for studying, e.g., lag phase, a more mechanistic cell division model would be more appropriate\(^2\).

**Population-level models vs individual-based models**

Population-level models (PLMs) directly describe the changes of the populations\(^3,4\). This can be achieved with differential equations if time is considered to be continuous or difference equations if time is considered to be discrete, e.g., if the population is stepped from generation to generation, or from year to year. PLMs are typically simple models that use a mass action approach for modelling interactions between different species. The mass action approach was adopted from kinetics of chemical reactions, where the probability of two molecules colliding and interacting is proportional to the concentration of each molecular species\(^5\). PLMs may consider the dynamics of resources explicitly, or assume a density dependence of growth rate such as logistic growth which implicitly considers resource depletion at higher population density\(^3,4\). The main advantages of PLMs are that they are relatively easy to describe and analyse, and they require less knowledge and data as they have fewer parameters. Their main purpose is to discover ‘general principles’ or concepts, as details are avoided. They have therefore been classified as ‘strategic’, ‘demonstration’ or ‘conceptual’ models\(^6\). Due to their general nature, they are less suitable to predict specific populations in specific ecosystems as would be desired for ecosystem management\(^6,7\).

PLMs are often based on ordinary differential equations (ODEs) or partial differential equations (PDEs); these are closely related, facilitating the exploration of corresponding ODE and PDE models\(^3,4\). ODEs assume homogenous space and are therefore appropriate for well-mixed systems such as chemostats or batch cultures. However, a non-uniform system may be represented reasonably well as consisting of different compartments. Then, a different set of ODEs for each compartment, and exchange between compartments, can approximate spatial structure. For example, a predator-prey model could have two types of habitat: one with food and predator, and one that is a refuge for the prey\(^3,4\).

PDEs are mostly used for fully spatially explicit models\(^3,4\). The effects of spatial structure can be inferred from comparing corresponding ODEs and PDEs. These are both continuum models, where time, space and species densities are continuous, rather than discrete, variables. As a result, populations may become infinitesimally small without becoming extinct. The similar difference equations are discrete in time and either discrete in population size or continuous in population density. If they are discrete in population size, extinction occurs more readily. As a consequence of using discrete time, population responses have a built-in delay, i.e. poor weather in one year affects the population size only in the next year. This delay renders dynamics less stable\(^3,4\).
Whilst PLMs usually neglect population heterogeneity, they can incorporate population structure in two ways\textsuperscript{4}. One is to separate the population into multiple age classes or life-cycle stages and describe the rate of change for each class by a separate ODE, e.g., ‘graduation’ from one age class to the next would be based on the growth rate. Another is to use a PDE to represent the population structure, e.g., age or size structure, as a continuum\textsuperscript{4}. Unavoidably, the more complex a PLM becomes, the more it loses its advantage over a corresponding IBM of being simpler and more tractable mathematically\textsuperscript{5}.

Individual-based models (IBMs), in contrast to PLMs, do not describe changes on the population level at all: they only describe the activities and properties of individuals, how they change the environment, and how they respond to the environment\textsuperscript{8–12}. Changes on the population level emerge automatically from all the interactions between individuals and the environment. Therefore, IBMs are classified as bottom-up models, describing the lower level to predict the higher level. For the same reason, PLMs are classified as top-down models. While PLMs can be made more and more complex to include population and spatial structure, thereby coming closer to IBMs, they remain top-down, describing the changes on the population level, rather than directly describing individuals like IBMs.

However, please note that this characterisation of various approaches simplifies in order to emphasize what is typical; in fact, there is a range of complexities for each type of model. For example, IBMs do not need to have spatial structure and can be quite simple to reveal general principles, while PLMs can be applied to particular systems and then become quite complex.

**Good modelling practices in IBM**

As mentioned above, an advantage of PLMs is that they are typically simpler to describe, understand and review. In order to facilitate the description of complex IBMs, the ‘ODD’ (Overview, Design concepts, and Details) protocol\textsuperscript{13} for systematic and complete description of IBMs has found widespread use. ODD is similar in purpose to MIRIAM\textsuperscript{14} (Minimum Information Requested In the Annotation of biochemical Models), but more standardization would benefit the IBM field just like the standards established for model exchange and description in systems biology have been highly beneficial. For example, the Systems Biology Ontology (SBO\textsuperscript{15}) provides an unambiguous vocabulary for model description and the Systems Biology Markup Language (SBML\textsuperscript{16}) enables the exchange of completely and unequivocally specified models.

Because IBMs map individual behaviour to population dynamics, they can bridge these two scales and use data on both levels: observations of individuals can be used as input into the model and observations of population dynamics can be compared with model output. Alternatively, individual behaviours can be inferred from comparing the kinds of population dynamics and patterns an assumed individual behaviour would produce with those dynamics and patterns observed, over a range of conditions. As mentioned in Box 2 (“Limitations of IBMs”), this has been called pattern-oriented modelling\textsuperscript{17}.

Individuals in IBMs are always discrete, but they may be either cells in a spatial grid (lattice) or particles in continuous space. In the first case, the IBM can also be called a cellular automaton (CA) where updating of lattice elements, diffusion of metabolites, cell division and movements are all specified by rules\textsuperscript{18,19}. Individuals typically occupy a single grid element and can move to a
neighbouring grid element only in certain directions (at angles that are multiples of 45° or 90°), which can lead to spatial artefacts\(^\text{30}\). A common rule for cell division is this: if a threshold mass is reached, divide into equal daughter cells. One daughter cell picks one of the free neighbouring cells at random, or if there are none free, pushes a random neighbouring cell away, which then moves according to the same rules\(^\text{21–23}\). To avoid artefacts, lattice elements should be updated in random order\(^\text{24}\).

Modelling individuals as particles with real size in a continuous space facilitates physically correct modelling of mechanical interactions between cells; these may be collisions or pushing away of cells that have encroached on one another due to motility or expansion of cellular volumes\(^\text{25–27}\). Such models can also be called particle-based models, but note that IBMs are only that subset of particle-based models where the particles may differ and have adaptive behaviour. Whether based on CAs or particles, a useful feature of IBMs is that behaviour can be described using simple rules and "if statements" that are not easily captured with differential equations\(^\text{8–12}\). For example, cell division is commonly triggered when a cell size threshold has been exceeded\(^\text{28}\).

Individuals in IBMs are autonomous agents that have their own state and carry out activities according to their state and in response to the environment. Hence, individual-based models are often called agent-based models. However, the term agent is more general as an agent does not have to be an individual. Agents can cover many scales, from molecular entities, cells, individual organisms, to social groups of organisms such as families, or larger social or economic organizations\(^\text{12}\).

Since IBMs explicitly simulate individuals, they can simulate population heterogeneity in a straightforward manner. As mentioned in Box 2 ("Limitations of IBMs"), IBMs of systems with a very high number of individuals generally do not explicitly simulate all microbial cells, but representative ones called “super-individuals”\(^\text{29,30}\). So, even in IBMs, there may be some lumping. Therefore, in terms of population heterogeneity, there is no hard distinction between PLMs and IBMs: The resolution increases smoothly from PLMs to super-individual IBMs to true IBMs. However, the two approaches are still fundamentally different in that the PLM describes the behaviour of the population and that the IBM describes the behaviour of individuals\(^\text{29,30}\).

**Modelling growth kinetics as an example.** This difference between PLMs and IBMs can be illustrated with the example of microbial growth, which is fundamental for any modelling of population dynamics. Growth kinetics are non-linear, and this has important consequences. In Droop kinetics, a commonly used model for growth of phytoplankton, the specific growth rate depends on the cell’s internal content of the limiting nutrient\(^\text{31}\). If internal nutrient contents vary between individuals, as observed in samples from the environment, the sum of the growth rates of all individuals will be different from the growth rate of a population with an average nutrient content\(^\text{31}\). This is an example of a well-known mathematical theorem, known as Jensen’s Inequality, that the average of a non-linear response to some heterogeneous input is different from the response to the average input\(^\text{32}\).

In Monod kinetics, the standard model for growth of heterotrophic bacteria, the specific growth rate does not depend on the internal nutrient content of the individual, but on the substrate concentration in the environment\(^\text{33}\). Thus, it could be modelled with PLMs or IBMs, depending on
the purpose of the model, e.g., whether other effects of individuality are to be considered or not. IBMs would be more appropriate for the purpose of modelling growth if the Monod kinetic parameters (maximum specific growth rate and substrate affinity) would differ between individuals. Such individual differences could be due to variation in expression of genes for uptake and metabolic enzymes between cells. Variation in maximal specific growth rates have been observed, most notably in populations with non-growing persister cells, but variation in substrate affinity between different cells has, to our knowledge, not been investigated. This could be studied in microfluidic single-cell chemostats. If individuals had different Monod kinetics, the kinetics of the population, which could be inferred with an IBM summing the rates of the individuals, would deviate from Monod kinetics. However, this would be difficult to observe in large populations, especially as individual growth rates fluctuate over time and faster growing lineages would become more frequent in the population over time and so come to dominate the population kinetics.

Models of intracellular dynamics. Intracellular dynamics, such as metabolism or gene regulation, can be integrated into IBMs, since IBMs have that flexibility of describing the activities of individuals by any means available to a programmer: from simple rules to complex, computationally expensive submodels. Focussing here on metabolism, there are two main ways in which metabolic fluxes can be predicted: dynamic kinetic models and steady state flux-balance models. Ideally, one would like to be able to use a dynamic kinetic model and write down the kinetic equations for all enzymes in a cell and then simulate the resulting fluxes through the metabolic network, from which growth rates could be predicted. The advantage of such dynamic kinetic models is that they can simulate the effect of changes in metabolite or enzyme concentrations, or in regulation of enzyme activity. However, this is not feasible for a genome-wide metabolic network, as the kinetics are only known for a limited number of enzymes from a limited number of species and often not under physiological conditions. Therefore, one usually either neglects large parts of the metabolic network or represents those parts as a stoichiometric model, and focusses instead on energy metabolism, where more is known.

For most species, even the kinetics of catabolic enzymes are not known sufficiently to use dynamic kinetic models. Or, one wants to include less well studied enzymes. Then, genome-wide flux-balance models, also known as constraint-based models, can be used instead because they only require knowledge of the list of enzyme reactions coded for by the genome and the stoichiometries of these reactions. However, the reaction rates can only be calculated when the equations are simplified by assuming that the system is in steady state, i.e., that the concentrations of the metabolites do not change with time. This is a reasonable assumption during exponential growth. Then, the distribution of fluxes (reaction rates) through the metabolic network that fulfil the stoichiometry can be calculated. To narrow down the space of possible solutions for these flux distributions, one uses constraints, the more the better. For example, using experimentally measured fluxes, placing upper bounds on reaction rates, using thermodynamics, or using gene expression data to remove reactions catalysed by those enzymes that are not produced under given conditions. To obtain a unique solution for the flux distribution within the narrowed down solution space, the one flux distribution that is optimal for the growth of the cell is picked.
Commonly, the objective or goal function for this optimization is to maximize biomass production (growth yield), although the choice of objective function can be debated. In conclusion, as all modelling approaches have their advantages and disadvantages, it is better to use a combination of approaches, as is true for experimental approaches. Therefore, the best option is to combine several modelling and experimental approaches.
References


SUPPLEMENTARY INFORMATION


