Evolutionary Applications
Open Access
Evolutionary approaches to environmental, biomedical and socio-economic issues

A seven-step guide to spatial, agent-based modelling of tumour evolution

Journal:	*Evolutionary Applications*
Manuscript ID	EVA-2024-024-RS.R1
Wiley - Manuscript type:	Review
Keywords:	Disease Biology, Evolutionary Medicine, Population Genetics - Theoretical, Development and Evolution
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
main.tex	

A SEVEN-STEP GUIDE TO SPATIAL, AGENT-BASED MODELLING OF TUMOUR EVOLUTION

Blair Colyer[1], Maciej Bak[1], David Basanta[2], and Robert Noble*[1]

[1]Department of Mathematics, City, University of London, London, UK
[2]Department of Integrated Mathematical Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA
*robert.noble@city.ac.uk

Author Contributions

Robert Noble conceived the idea for the review; Blair Colyer gathered the review material; Blair Colyer and Robert Noble led the writing of the manuscript with contributions from David Basanta and Maciej Bak. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgments

We are grateful to Jill Gallaher, Rafael Bravo and Mark Robertson-Tessi for helpful conversations about model implementation, and to Martín López-García and Grant Lythe for alerting us to potential uses of phase-type distributions.

ABSTRACT

Spatial agent-based models are frequently used to investigate the evolution of solid tumours subject to localised cell-cell interactions and microenvironmental heterogeneity. As spatial genomic, transcriptomic and proteomic technologies gain traction, spatial computational models are predicted to become ever more necessary for making sense of complex clinical and experimental data sets, for predicting clinical outcomes, and for optimising treatment strategies. Here we present a non-technical step by step guide to developing such a model from first principles. Stressing the importance of tailoring the model structure to that of the biological system, we describe methods of increasing complexity, from the basic Eden growth model up to off-lattice simulations with diffusible factors. We examine choices that unavoidably arise in model design, such as implementation, parameterisation, visualisation, and reproducibility. Each topic is illustrated with examples drawn from recent research studies and state of the art modelling platforms. We emphasise the benefits of simpler models that aim to match the complexity of the phenomena of interest, rather than that of the entire biological system. Our guide is aimed at both aspiring modellers and other biologists and oncologists who wish to understand the assumptions and limitations of the models on which major cancer studies now so often depend.

Introduction

Cancer initiation, progression, and treatment responses are Darwinian evolutionary processes [1, 2] that can be investigated using a wide range of mathematical and computational methods. Examples include evolutionary game theory [3, 4], branching processes [5, 6], and Moran processes [7, 8]. Yet while many tools have yielded important insights into cancer evolution, the study of spatial aspects – especially important in carcinomas, constituting the majority of humans cancers – often necessitates a spatially explicit approach, such as a spatial agent-based model.

An agent-based (or individual-based) model is a computational model of a system made up of autonomous, interacting "agents". Spatial agent-based models (SABMs) have long been used to study the evolution of spatially structured communities because they can reveal how the processes of selection, drift, and gene flow depend on localised interactions among agents (typically individual organisms) or between agents and their spatially varying environment. As new technologies generate better spatial tumour data, SABMs are proving ever more useful in oncology. Typical applications include understanding tumour development, inferring the effects of driver mutations, and predicting treatment outcomes. For example in recent studies, Aif *et al.* [9] used an SABM to investigate the evolutionary rescue of drug-resistant tumour subclones; Saha *et al.* [10] used an SABM to investigate adaptive cancer therapy; and Bull and Byrne [11] used an SABM to simulate interactions between macrophages and tumour cells.

To support this burgeoning research field, here we present a seven-step guide to designing and implementing spatial agent-based models in which the agents are locally-interacting tumour cells or cell subpopulations. Starting from the

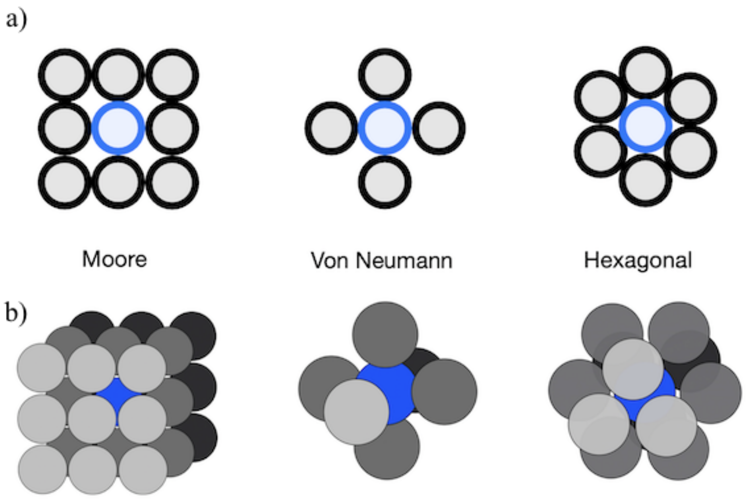

Figure 1: Some common neighbourhoods that govern the update rules for cellular automata and other agent-based models in two dimensions (**a**) and three dimensions (**b**). A focal agent (cell) is shown in blue and its neighbourhood sites in grey.

simplest cellular automata, we discuss options for adding greater complexity and biological realism, such as multi-level spatial structure and environmental heterogeneity. Based on our extensive experience of developing and using SABMs [12, 13, 14, 15], we cover practical issues such as event scheduling, visualisation, and how to use SABMs to infer parameter values from experimental or clinical data. Each topic is illustrated with examples from our own demon-warlock modelling framework [14, 12], other state of the art modelling platforms, and studies that have used SABMs in cancer research. Whereas our focus is on tumour evolution, much of our advice applies equally to similar modelling methods used to study bacterial colonies, invasive species, and organismal development. The guide is designed to be accessible for biologists and clinicians without specialist mathematical knowledge.

1 Spatial structure

Spatial structure determines the evolutionary balance between selection and drift, the nature of gene flow between subpopulations, and the strength of ecological interactions. When a model fails to accurately represent the spatial structure of a biological system, the model's predictions and inferences for that system may be highly unreliable [12, 16]. It follows that the parameters of spatial structure – such as the size of locally interacting cell communities and the manner of cell dispersal – should be accorded the same importance as evolutionary parameters in model design. Notwithstanding the trade-off between model simplicity and realism, spatial structure parameters should, as far as possible, be derived or inferred from empirical data.

1.1 Stochastic cellular automata

Many of the simplest spatial agent-based models are cellular automata. A cellular automaton is a model that plays out on a grid of sites in one or more dimensions. Each site is associated with one of a set of at least two possible states.

Each site also belongs to a subset of sites called a neighbourhood, of which some examples are shown in Figure 1. For example, the von Neumann neighbourhood in two dimensions contains the nearest sites in the cardinal directions (up, down, left and right). A cellular automaton sequentially updates itself according to a set of rules. The update rules for a given site depend on its own current state and the states of the sites in its neighbourhood.

Whereas the update rules of many cellular automata are deterministic [17], probabilistic rules are more appropriate for modelling stochastic processes such as biological evolution. A stochastic cellular automaton is equivalent to a collection of locally interacting Markov chains, which means that each event is chosen according to probabilities that depend only on the current model state, not any of its previous states.

In biological terms, each state corresponds to a type of cancer cell or some other entity (such an immune cell or part of the extracelluar matrix). Generally we will assume that the focal agents in our models are cancer cells and we will use the terms "agent" and "cell" interchangeably where appropriate. A cellular automaton permits a cell's event probabilities (for example, its division, death, and dispersal rates) to depend on the number of neighbouring cells. This allows us to account for crowding or Allee effects, such that birth, death or dispersal rates depend on the local or global population size. Event rates can also vary according to the types of the neighbouring cells, for example to simulate cell competition or immune predation.

Models of asynchronous processes, such as cell division in a tumour, typically use asynchronous updating, meaning that only one or a small number of sites are modified per update [18]. In addition to being more realistic, asynchronous updating is often necessary to prevent conflicts. For instance, if two cells are attempting to divide but only one space is available for the two potential daughter cells then one must take priority.

1.2 The Eden growth model

Among the simplest stochastic cellular automata is the Eden growth model. This model is typically implemented on a two- or three-dimensional regular square grid with only two possible states: unoccupied (S_0) and occupied (S_1). With each iteration, the update rule causes a site in the neighbourhood of an S_1 site to switch from S_0 to S_1. In this way new S_1 sites (cells) are added to the surface of a cluster. The Eden growth model on an n-dimensional grid self-organises to resemble an n-dimensional ball with a non-trivial surface. The growth curve of the S_1 population approaches a polynomial of degree n [19].

The three most popular options for the Eden growth model update rule can be labelled alphabetically:

- **A**vailable site-focussed: Choose at random an S_0 site in the neighbourhood of an S_1 site, and switch it from S_0 to S_1.

- **B**ond-focussed: Choose at random an S_1 site with a probability proportional to the number of S_0 sites in its neighbourhood, and then randomly choose an S_0 neighbour and switch it to S_1.

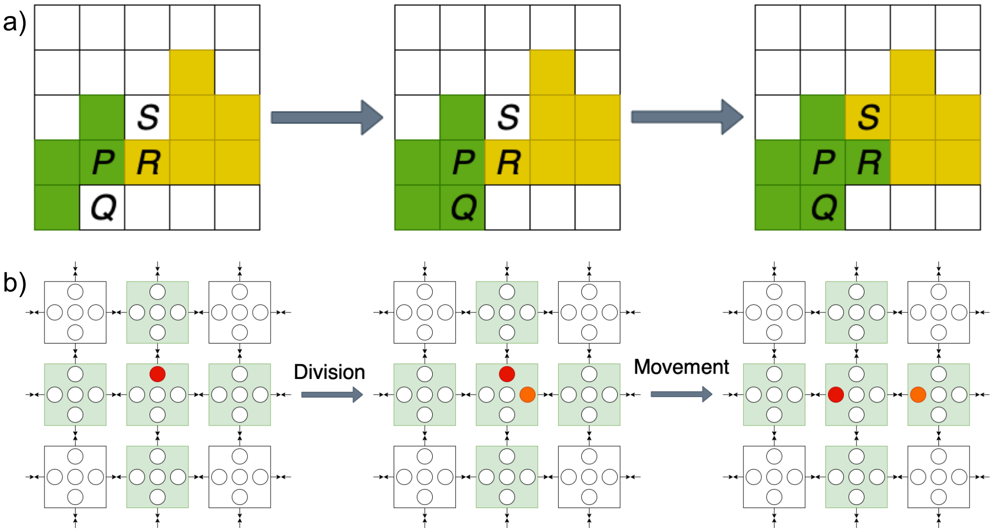

Figure 2: **a**) An illustration of the CPM. Here, two cells, shaded in green and yellow, each occupy multiple sites on a grid. In the leftmost panel, we see the model's initial configuration; in the middle panel, the state of site P is copied onto site Q, and the green cell grows; in the rightmost panel, the state of site P is copied onto site R, which is initially occupied by the yellow cell, thus deforming the yellow cell and budging it into site S. **b**) An illustration of the LGCA. Grid sites shaded in green represent those which may influence the focal cell node, shaded in red. The leftmost panel represents the initial configuration; the middle panel shows a cell dividing into free space on its grid site, with the new cell shaded in orange; the rightmost panel demonstrates how cells might move within a grid site or between grid sites: the red cell has changed direction, and the orange cell has moved from one site to another.

- Cell-focussed: Choose at random an S_1 site with at least one S_0 site in its neighbourhood, and then randomly choose an S_0 neighbour and switch it to S_1.

Although these update rules result in similar large-scale patterns, they generate cluster surfaces with different microscopic properties. Indentations in the model surface are more likely to be filled, and spikes are less likely to form, under option C than under option B, and under option B than under option A. Hence option C generates the smoothest surface and option A the roughest [20].

Variants of the Eden growth model have been used to investigate the evolution of paediatric glioma [21], colon cancer [22], hepatocellular carcinoma [23, 24], clear cell renal cell carcinoma [25] and non-small cell lung cancer [26]. Many studies use a variant that includes stochastic cell death. By opening up spaces for cell division, cell deaths increase clonal mixing within the tumour and facilitate selection [23].

1.3 Other grid-based stochastic cellular automata

Other stochastic cellular automata can be more appropriate than the Eden growth model for modelling systems in which state changes are not confined to the surface. Spatial branching processes are similar to Eden growth models except that if a dividing cell has no space to divide then it can create space by budging other cells. An intermediate model can be created by stipulating that only nearby cells can be budged, so as to simulate physical constraints on cell division. Chkhaidze *et al.* [27] recently used such a model to investigate how spatially constrained tumour growth alters signatures of clonal selection and genetic drift in cancer genomic data. Good practice is to implement budging along an approximately straight line between the dividing cell and the nearest empty site. If budging is instead restricted to the cardinal directions or the cardinal and intercardinal directions then the simulated tumour will self-organise into an approximate square or octahedron, rather than a more biologically plausible disc or ball.

Another option is to allow dividing cells to replace, rather than displace, their neighbours. In the voter model, the update rule is such that, with a certain probability, a randomly selected site copies the state of a neighbouring site. Biasses can be introduced by setting unequal copying probabilities, corresponding to differences in cell fitness. Simple (linear) voter models satisfy a convenient property called coalescing duality, which means that their typical behaviour can be explained through mathematical analysis [28]. In a pioneering 1972 study, Williams and Bjerknes [29] used a biassed voter model to simulate the spread of skin cancer through the basal epithelial layer.

The cellular Potts model (CPM), also known as the Glazier-Graner-Hogeweg model [30, 31], more explicitly simulates physical interactions among cells and between cells and their microenvironment. The model takes place on a lattice and each cell is represented by multiple lattice sites (as opposed to only one lattice site, as in previously discussed models), corresponding to the cell's volume (Figure 2a). Cells are deformable and can adhere to one another or to surrounding empty sites (which might represent extracellular matrix or growth medium). Hamiltonian mechanics describe the overall energy of the system depending on adhesion forces and resistance to changes in cell volume. A random lattice site is chosen at each time step and its state is copied to a random neighbouring site. If the new configuration has lower energy than the previous configuration then the change is always accepted; otherwise, the probability of accepting the change depends on the Boltzmann temperature. The CPM has been used in numerous cancer studies, such as for simulating tumour growth, invasion and evolution [32], or for investigating how cell compressibility, motility and contact inhibition shape tumour cell clusters [33]. The CompuCell3D modelling environment compucell provides an efficient, flexible CPM implementation.

The biological lattice gas cellular automaton [34] excels instead at modelling cellular movement, and especially collective migration, in a simple, computationally efficient, and physically correct fashion. The model must play out on a square or hexagonal lattice in 2 dimensions, or a cubic, dodecahedral or icosahedral lattice in 3 dimensions. States incorporate cell velocities. For instance, consider a 2-dimensional square lattice in which each site contains 5 nodes: one for each directional velocity and a resting node at the centre (Figure 2b). A cell occupying any one of these nodes can divide into other nodes on the same site. A cell can also reorient itself by moving between nodes on the same site, and can move between sites according to its velocity, provided there is space to do so. This model has been used, for example, to give insights into breast cancer invasion plasticity [35].

1.4 Multi-level spatial structures

An important limitation of all the aforementioned cellular automata is that their uniform spatial structures are inconsistent with the biology of many tumour types. Various common cancers have glandular structures and grow via individual cells or small cell clusters invading neighbouring tissue [36, 37]. Colorectal adenomas are also glandular but grow through gland fission [38].

Inspired by classical population genetics models [39], a simple, conventional way to account for multi-level spatial structure in tumours is to assign cells to local subpopulations, called demes, located on a regular grid. Thus each grid site is allowed to contain not only one but dozens, hundreds, or thousands of cells. The subpopulation size per deme is prevented from exceeding a certain threshold – known as the deme's carrying capacity – by decreasing cell division rates or increasing death rates as the subpopulation size grows.

Deme-based models allow for more complicated modes of cell dispersal. As in the voter model, cells can be assigned some probability of invading neighbouring demes, either individually or in clusters. The dispersal probability can also be made to depend on the population of the deme being invaded, so that cells disperse more easily in less densely populated regions near the tumour periphery. Alternatively, each occupied deme can be assigned a probability of undergoing fission, resulting in some of its cells being moved to an unoccupied neighbouring deme. Depending on the degree of budging allowed, the deme-level dynamics of the fission model can resemble an Eden growth model (no budging of demes) or a spatial branching process (unlimited budging). Deme-based models additionally allow for the explicit simulation of tissue invasion, such that a tumour can grow only via its cells invading demes that are initially filled with normal cells [12].

1.5 Aggregating agents

If the within-deme subpopulations can be assumed to be well-mixed then cells that belong to the same deme and have the same phenotype and genotype can be modelled collectively, rather than as individual agents. This model design not only improves computational efficiency but can also facilitate mathematical analysis. For example, when cells disperse by invading neighbouring demes, the model can be designed so that the dynamics are approximately equivalent to the well understood spatial Moran process [12]. Cells can be randomly selected within a deme by sampling from a hypergeometric distribution.

Even greater efficiency can be realised by not modelling inter-deme dynamics at all, and simply making the demes themselves the model agents [40, 41]. Although such coarse-graining enables the simulation of much larger tumours, it comes at the cost of reduced precision. Care should be taken in translating between mutation rates per cell and effective mutation rates per deme.

1.6 Off-lattice models

Instead of confining agents to a regular grid, we might instead locate them in continuous space. This structure is potentially more realistic but also entails more parameters, more decisions to be made, and typically higher computational costs [42]. To prevent multiple cells occupying the same space and to maintain tumour integrity, we now must model the movement of cells in response to physical forces such as cellular adhesion and repulsion [43]. We may also choose to model directed movement under the influence of diffusible factors (hapotaxis).

There are several practical ways to prevent cells overlapping in an off-lattice model, depending on how the agents are implemented. Suppose we have spherical cells, each with fixed radius r. We can then specify that when, as a result of cell division or movement, the distance between two cells' centres is less than $2r$, both cells will simply be pushed in opposite directions. Alternatively, to account for cell deformation, we might implement repulsion only when the distance between cell centres falls below some threshold value smaller than $2r$ [44]. Some modelling platforms achieve greater realism and tractability by implementing adhesion and repulsion forces using functions rooted in physics, which are beyond the scope of this guide (see documentation cited in the appendix).

2 Mutation

Having chosen an appropriate spatial structure, we next will decide which cell phenotypes and genotypes to include in our state space, and how to model mutations between these states. As ever, the goal is to balance model simplicity, realism, and computational demands.

2.1 Defining phenotypes

A good part of the difficulty in designing a useful model stems from the fact that much of the experimental data gathered by cancer biologists focusses on genetic mutations while the rules that govern the behaviour of the agents in an SABM assume an understanding of the key cancer phenotypes. The most basic actions a tumour cell might perform at any given time step are apoptosis/death, proliferation, and motility. These are often considered as simple probabilistic events and often modelled in a exclusionary manner, so that if a cell is moving then it is neither proliferating nor dying. The required probabilities can either be taken directly from experimental data (which is often hard to measure *in vivo* and unrealistic *in vitro*) or calibrated with *in vivo* pre-clinical models.

Using hard-coded rules to model the phenotype of a tumor cell, while relatively simple, does not capture the flexibility shown by biological cells in the mapping between genotype and phenotype. Gerlee and colleagues have instead proposed capturing some of the complexity of this mapping by embedding neural networks inside each agent, so that the phenotype emerges in a non-linear way as a result of the agent's state and the different microenvironmental inputs to which the agent is receptive [45].

2.2 Trait evolution versus population (epi)genetic models

Once phenotypes have been defined, the next step is to determine how these phenotypes will change as a result of mutations. One option is to model mutations as phenotypic switches. Many studies consider models with only two possible tumour cell states – mutated and unmutated – which differ in fitness [40], degree of drug resistance [46], or some other trait. Grow-or-go models assume that cells can reversibly switch between predominantly migratory and

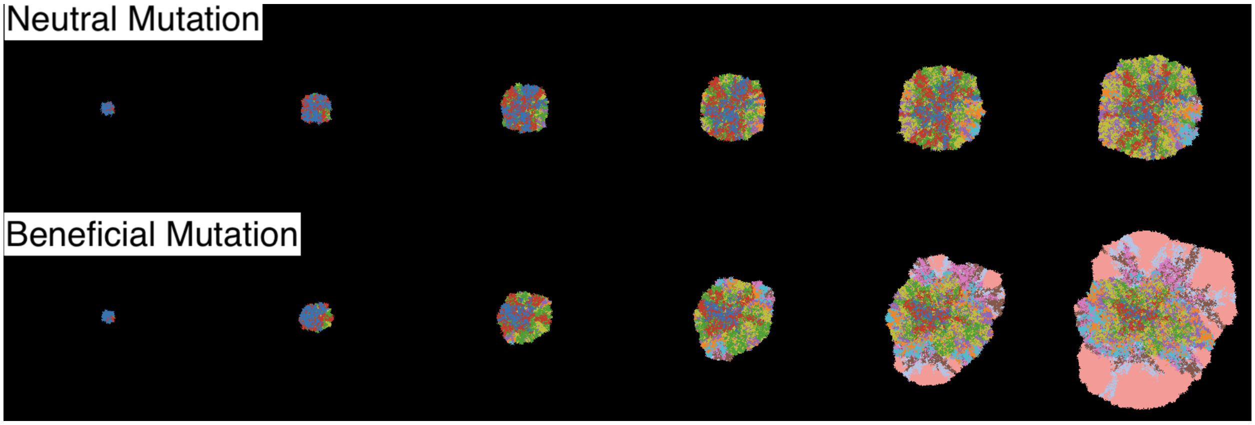

Figure 3: The result of running an Eden growth model with nearly neutral mutations (top) and beneficial mutations (bottom). Model produced in HAL using some in-built examples as a skeleton for the code. [51].

predominantly proliferative phenotypes [47]. Other models examine the evolution of continuous traits, such as levels of glycolysis and acid production [48].

If we are more interested in clonal dynamics then we can explicitly track changes to the (epi)genome. These mutations are conventionally assigned to three groups according to how they affect cell fitness: driver mutations (which increase cell fitness), passenger mutations (no effect), and deleterious mutations (negative effect). For simplicity, most studies assume an infinite sites model [49], such that no two mutations can occur at the same site. Finite sites models must be parameterised based on observed mutation frequencies [50].

2.3 Example: The Eden growth model with mutation

We can convert an Eden model into an evolutionary model by implementing mutation. The grid and neighbourhood are defined as before but now we have multiple cell states $S_1, S_2, S_3, \dots$ and mutation rates between each pair of distinct cell states. A simple option, assuming infinite sites, is to set all mutation rates to be zero except in the case of S_i to $S_i + 1$ for all $i \geq 0$, so that every S_i cell has exactly i mutations. Let us assume that all these mutations are drivers and their effects combine multiplicatively, such that each mutation increases the division rate by a factor of $1+s$, with $s \geq 0$. Assume also that mutations occur only at the time of cell division, and the number of new mutations per daughter cell is Poisson distributed. We then arrive at a reasonable toy model of spatial tumour evolution that can be implemented in not much more than 100 lines of code, as we illustrate with an R script [52]. Figure 3 shows results of implementing a similar model in the HAL platform [51].

2.4 Distributions of fitness effects

Modelling the evolution of a quantitative trait, such as cell division or death rate, leads to further design decisions. As in our toy model, it can be wiser to draw mutation fitness effects from a probability distribution instead of setting them all equal. To see why, consider a model of an expanding tumour that, in the absence of mutation, has radius growth rate c_0, and in which the spread of mutants is not confined to the periphery (for example, a biassed voter model).

When a new mutant arises within the wildtype population, its long-term fate, in the absence of further mutation, will be sensitive to its radius growth rate, c_1. If $c_1 < c_0$ then the mutant will remain forever rare; if $c_1 > c_0$ then the mutant is bound to take over the entire tumour; if $c_1 = c_0$ then the mutant will become relatively more abundant over time without ever fully replacing the wildtype. Randomising the fitness effect randomises c_1 and so randomises mutant fates. Our demon-warlock framework draws each selection coefficient (relative increase in cell division rate) from an exponential distribution.

Strictly multiplicative fitness is best avoided in all but the smallest-scale models as it can lead to unrealistically high fitness values. This is especially problematic if mutation is implemented at the point of cell division, which creates a feedback loop in which lineage fitness grows at an ever increasing rate. A simple solution implemented in our demon-warlock framework is diminishing returns epistasis. When the selection coefficient of a driver mutation is s, instead of multiplying the division rate by $1 + s$, we instead multiply by $1 + s(1 - b/b_{max})$, where b is the previous division rate and b_{max} is an upper bound.

3 Event scheduling

The next step is to consider how to implement cell events algorithmically. Event scheduling can be the most important factor in determining computational efficiency, especially in simpler grid-based models. The optimal choice strikes a balance between efficiency, simplicity, and biological realism.

3.1 Gillespie's algorithm

The Gillespie Stochastic Simulation Algorithm [53] is an especially simple and popular solution to event scheduling. Event rates are assumed to depend only on the current state of the model and the time between events is exponentially distributed (as in a Poisson process), such that two events cannot occur simultaneously. The steps of the algorithm are as follows:

1. Initialise the system.
2. Set event rates (birth rates, death rates, dispersal rates, etc.).
3. Randomly determine the next event such that $\mathbb{P}(event = E) = rate(E)/\Sigma(rates)$
4. Implement the chosen event.
5. Advance the timer by $\delta t \sim \text{Exp}(1/\Sigma(rates))$
6. Repeat from step 2 until a stop condition is reached.

This algorithm is more efficient than the event timer approach (see below) and is very easy to implement. In statistical terms, the simulated sequence of events corresponds to a trajectory of a set of stochastic differential equations, called the master equations. This means we have a good mathematical understanding of how the algorithm behaves.

Our toy Eden growth model [52] provides an example implementation of Gillespie's algorithm. This model further improves computational efficiency by keeping track of the cells that have space to divide, so that the next dividing cell can be chosen from among this subset (which in n dimensions scales with the radius to the power of $n - 1$) rather than from the entire cell population (which scales with the radius to the power of n). The drawback is that cells without space to divide never undergo mutation, which may be an unjustifiable assumption in a serious research model.

Modifications of Gillespie's algorithm, such as tau leaping [54], are even faster but less accurate. Tau leaping allows multiple events to occur simultaneously, which may be problematic in a spatial model if the events affect multiple sites in close proximity (for example, if two cells are chosen to divide into the same empty site). Moreover, tau leaping improves performance only when the system is dominated by a small number of large, homogeneous subpopulations, which is typically not the case in SABMs.

3.2 Gillespie's algorithm with phase-type distributions

A shortcoming of the Gillespie algorithm is that some events, such as cell division, are not true Poisson processes with exponentially distributed waiting times. In effect, the Gillespie algorithm permits arbitrarily short cell cycles. Some cells may divide several times while, in the same period, others with identical division rates fail to divide at all.

One way to achieve more realistic cell cycle periods without sacrificing very much computational efficiency is to use a phase-type probability distribution. Whereas an exponential distribution models the time until the next event in a Poisson process, a phase-type distribution models the time taken for an entire sequence of events, which may occur at different rates.

In practical terms, this entails executing the Gillespie algorithm as above, except that when a cell is selected for division, it doesn't necessarily divide immediately, but instead changes its position in the cell cycle. Given a target probability distribution for cell cycle periods, we can use an algorithm to choose transition rates such that the resulting phase-type distribution has the same mean, variance, and skew as the target [55]. For example, suppose that all cells begin in division state 0. When a cell is selected (according to a state-dependent probability), its state is updated from 0 to 1, 1 to 2, or 2 to 3. When a state 3 cell is selected it divides and both progeny are reset to state 0 [56]. The method's greater realism comes at the cost of additional memory demands and longer execution time, compared to the basic Gillespie algorithm.

3.3 Random sampling with binary trees

When we have more than a handful of events to choose from it will be much more efficient to implement event selection using a binary tree. Suppose, for example, that we have four possible events with rates p_1, p_2, p_3 and p_4, where $p_1 \leq p_2 \leq p_3 \leq p_4$. If we store the rate sums $p_1 + p_2$, $p_3 + p_4$, and $p_1 + p_2 + p_3 + p_4$ then we can choose an event as follows. First we generate a random number r from a uniform distribution between 0 and $p_1 + p_2 + p_3 + p_4$, and we examine whether $r < p_1 + p_2$. Supposing r is greater than $p_1 + p_2$, we then test whether it is less than p_3. If

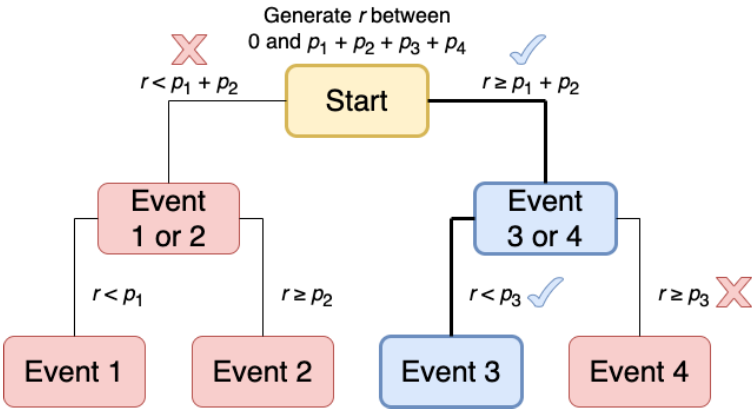

Figure 4: An example of using a binary tree to select an event (Event 3) from four options. Selected nodes are shown in blue.

so then we choose event 3; otherwise event 4. Effectively, we have traversed a binary tree, beginning at the root node associated with the sum of all event rates, and ending at a terminal node associated with a single event (Figure 4).

The binary tree method is efficient because both the number of steps needed to choose an event, and the number of nodes that need updating following a change in an event rate, grow only with the logarithm of the number of possible events. For example, we need only twenty steps to choose between one million possible events. As long as the cell population keeps growing, there is little benefit to pruning nodes and it is easy to ensure that the tree remains balanced. The rate sums together take up only as much computer memory as the individual rates. The main costs are in terms of code development time and code complexity. Binary trees require careful implementation and error checking to ensure that existing nodes are updated and, when required, new nodes are added after each model event. Our demon model implements binary trees and periodically recalculates event rate sums to prevent excessive accumulation of rounding errors.

3.4 Cell cycle timers

A less efficient alternative to using phase-type distributions is to draw inter-division times directly from a chosen probability distribution. This approach enables more precise tracking and adjustment of individual cell cycles. An algorithm used in recent studies [48, 46] is as follows:

1. Initially assign every cell i a countdown timer set to time t_i drawn from some probability distribution (dependent on the cell's phenotype).
2. Subtract δt from every countdown timer, where $\delta t \ll t_i$ for all i.
3. For all cells i, in random order:

   A. Implement cell death and dispersal events for i;

   B. If i is alive, has space to divide, and $t_i \leq 0$, then i divides;

   C. Assign each new cell a countdown timer, set to some random time dependent on the new cell's phenotype.

4. Repeat from step two until a stop condition is reached.

How much this approach reduces computational efficiency will depend on other aspects of the model. It is likely to be much slower than a well implemented Gillespie algorithm when applied to a simple grid-based model, due to the additional burdens of updating every cell (Step 2) and shuffling all the cells (Step 3) at each small time step. In an off-lattice model, where cells move much more frequently than they divide, and where a shuffling algorithm may already be required to randomise the order in which cell positions are updated, the cost of updating cell division state at the same time as position may be negligible.

4 Microenvironment

Whereas many SABM studies focus on the effects of spatial structure and cell-cell interactions, real tumours evolve in a complex microenvironment that varies over space and time. This tumour microenvironment, comprising both molecular elements, such as cytokines, and other (non-cancer) cells, constitutes the cancer ecosystem [57] – a key element of the selection process driving somatic evolution. Given a good rationale and sufficient parameterisation data, we may choose to extend our model by explicitly simulating microenvironmental factors in the form of agents (in the case of immune cells or stromal cells) or diffusible factors (such as oxygen and drugs). Permitting cancer cells to modify their selective environment creates potential for emergent complexity and niche construction [58, 59].

4.1 Hybrid cellular automata

Hybrid cellular automata (or HCA) have been used to model interactions between tumour cells and diffusible factors for more than twenty years. As described in a pioneering 2001 paper by Patel and colleagues [60], these models consist of two interdependent components: stochastic cell events, and deterministic reaction-diffusion partial differential equations. The latter component dictates how chemicals or other factors work their way through the system as they are consumed and processed by cells. Local concentrations of diffusible factors contribute to the cell update rules.

Typically we assume that diffusible factor concentrations rapidly re-equilibriate following changes in the configuration of cells. We can then numerically solve the equations to find the equilibrium concentrations either after every cell event or, trading some accuracy for greater efficiency, after a relatively small number of cell events have occurred. Suitable procedures for solving partial differential equations as initial value problems can readily be found in textbooks and software libraries. These range from simple but inefficient algorithms based on the classical Gauss-Seidel method, which require only a few dozen lines of code [61, 60, 15], to the highly sophisticated BioFVM solver [62], which is specifically optimised for hybrid SABMs. Several SABM platforms include their own methods for solving reaction-diffusion equations in two or three dimensions (see appendix).

4.2 Types of diffusible factor

To add biological realism, we might make cell division and death rates in our model depend on the local oxygen and glucose concentrations as these factors diffuse through the tumour from the surrounding medium (in very small tumours and tumour spheroids) or from point sources representing blood vessels (in larger, vascularised tumours). We might also modify dispersal rates so that cells follow oxygen or glucose gradients. Potential adverse factors include acid produced through tumour cell metabolism, and drugs that diffuse from blood vessels. Hybrid cellular automata are especially suitable when the supply of an influential factor is highly variable over space or time, such as in the case of intermittent drug treatment [63].

5 Parameterisation and inference

Although theoretical models can be valuable for generating hypotheses and providing proof of concept, if we want to apply an SABM to studying a particular biological system then we must ensure that its influential parameter values are set appropriately. Parameterisation should ideally be based on clinical or experimental data specific to the biological system of interest; otherwise values can be estimated from studies of similar systems or theoretical considerations (for instance, diffusion coefficients approximately correlate with molecular weight). Influential parameters might pertain to the effects of mutations, drugs, oxygen and glucose; rates of chemical supply, diffusion, consumption and decay; cell dispersal modes and rates; baseline cell death rates, crowding effects and the size of interacting cell communities. Since calibrating SABMs is often computationally demanding, high-performance computation may be required to generate the necessary resources to calibrate them properly.

5.1 Example: Hybrid cellular automaton for simulating a tumour spheroid

Bacevic and Noble *et al* [15] parameterised a HCA to mimic tumour spheroid evolution under drug treatment. In spheroids the limiting factor for cell survival and proliferation is oxygen. Other diffusible factors such as glucose were therefore omitted to simplify the model without compromising its usefulness. The oxygen concentration in the medium and oxygen diffusion rates were drawn from previous studies [64, 65, 66], as were the mathematical relationships between oxygen consumption rate, cell proliferation rate and local oxygen concentration [67, 68]. The different maximum proliferation rates of drug-sensitive and resistant cells, reflecting a fitness cost of resistance, were determined from new monolayer growth assays. Cells with insufficient oxygen supply were assumed to die.

Since oxygen effects alone fail to account for the extent of quiescence observed in tumour spheroids, Bacevic and Noble *et al* implemented crowding effects by permitting cell budging only within a specified radius. New monolayer growth assays revealed that the relationships between cell proliferation rate, death rate and drug dose could be well approximated with piecewise linear functions. The drug's impact on proliferation was further assumed to multiply the oxygen effect, consistent with prior observations [67]. Drug consumption was also modelled using Michaelis-Menten kinetics, with a diffusion rate chosen according to the drug's molecular weight and an appropriately low consumption rate. Thus parameterized, the SABM accurately predicted the outcomes of new tumour spheroid experiments [15].

5.2 Example: Hybrid cellular automaton of the bone ecosystem in cancer

Araujo and colleagues [69] developed a hybrid cellular automaton for which the goal was to capture the ecosystem of the bone. A crude approximation of this ecosystem includes the bone itself, the myeloid-derived cells such as osteoclasts that resorb bone, and the cells derived from messenchymal stem cells, such as osteoblasts, that deposit new bone. Each of these cell types can be modelled as discrete agents regulated by diffusible factors – such as TGF-β, RANK ligand, and other factors embedded in the bone matrix – described by partial differential equations. Parameterisation of the model is facilitated by the fact that non-cancerous cells have more predictable phenotypes, and the model's overall behaviour can be calibrated to ensure it recapitulates bone homeostasis. Araujo and colleagues thus studied how bone metastatic prostate cancer cells could infiltrate the bone ecosystem, take advantage of it, and grow [70]. They also investigated what prostate cancer cells in the primary tumour should be of concern to physicians, and why conventional treatments that fail to disrupt tumour-ecosystem interactions also fail to provide long-term cancer cures in bone metastatic prostate cancer [71].

5.3 Parameter inference

Unknown parameter values can be inferred by combining an SABM with a statistical method. This is, in fact, often the main objective of an SABM study. Approximate Bayesian computation is a popular approach that, in its simplest form, infers the value of a parameter θ as follows

1. From our data, calculate some summary statistic μ_{data};
2. Set $i = 1$;
3. Run the model using a candidate parameter value θ_i drawn from some prior distribution;
4. Calculate the summary statistic μ_i for the model output;
5. If the difference between μ_i and μ_{data} is less than a predefined tolerance then add θ_i to the posterior distribution;
6. Increment i;
7. If i is less than some threshold then repeat from step 3.

Although simple in principle, approximate Bayesian computation requires careful implementation. The accuracy and precision of inferences depend on the choices of prior distributions, summary statistics, and tolerances, as well as the number of iterations. Typically multiple parameter values cannot be precisely derived from prior data or models, in which case each should be assigned a vague (high variance) prior distribution. Tolerance values should be tuned such that neither too many nor too few candidate parameter values are accepted to the posterior distribution. Summary statistics should capture features of the system that provide useful information about the parameters of interest. A useful template is a 2010 study [72] in which Sottoriva and Tavaré inferred aspects of stem cell dynamics in the colonic crypt by combining a cellular Potts model with approximate Bayesian computation, using a summary statistic based on methylation patterns.

An alternative to this approach was recently outlined in [73], in which the authors describe a novel method utilising neural networks to reduce both tumour images and SABM simulations to low-dimensional points. The distance between these points acts as a quantitative measure of how the two differ. This enables direct comparison, and by using parameter fitting algorithms to minimise the distance between the two sets of points, parameters can be estimated directly from the images and the simulations.

5.4 Sensitivity analysis

Whatever the objective, an essential step in any modelling study is to examine, as far as is practical, how the results and conclusions depend on uncertain aspects of the model. A common approach is to run a large number of model variants with different combinations of plausible parameter values. Varying one parameter at a time can provide useful insight into which parameters have the greatest impact on model output, with the shortcoming that non-linear interactions between parameters are often neglected. A more sophisticated approach is to infer a multivariable "metamodel" function – a model of the model – that approximately describes how the model's parameters determine its outputs.

Since varying many parameters systematically on a continuous scale is infeasible, sampling methods such as Sobol sequencing [74] or Latin hypercube sampling [75] can be used to generate a set of near-randomly sampled combinations of parameter values. Both methods were used in a recent SABM study of cancer cell response to ATR-inhibitors [22]. A recent introductory review explains specifically how to apply these methods to cancer ABMs [76]. It is important to note that thorough sensitivity analysis involves varying not only parameter values but also mathematical relationships, aspects of spatial structure, and any other influential model components.

6 Visualisation

Having built and parameterised a model, we next require useful ways to visualise its output. Typical methods represent spatial information, multidimensional phenotypic information, or evolutionary dynamics. Representing all these aspects in a single image is generally impossible.

6.1 Spatial plots

A spatial plot represents the state of an SABM at a moment in time. Producing a two-dimensional grid plot of a two-dimensional on-lattice model is straightforward – we simply output the state of each site as a matrix of numbers and input this matrix into a bitmap (or raster) plotting function in R, Python, MATLAB, or similar software, using different colours to represent the different states (Figure 3). Our toy Eden growth model [52] provides an example

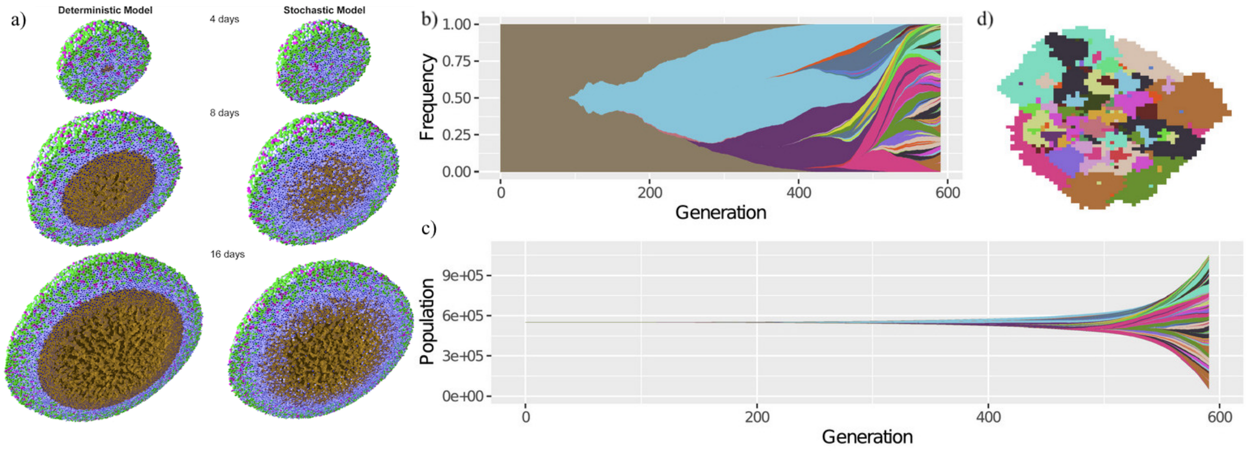

Figure 5: **a)** Plots of a 3D off-lattice ABM, produced in PhysiCell [77], showing a cross-section of model states of a hanging-drop spheroid growth simulation at different time points, using either a deterministic or a stochastic SABM. Cells are coloured according to cell cycle position. Cells in the K_1 cell cycle state are green, post-mitotic K_2 cells are magenta, quiescent cells are pale blue, apoptotic cells are red, and necrotic cells are brown. Cell nuclei are shown in dark blue. **b)** Muller plot showing phylogenies and phenotype frequencies over time. **c)** Fish plot showing phylogenies and phenotype population sizes over time. **d)** 2D grid plot corresponding to same simulation as the Muller and fish plots in previous panels, with the same colour scheme, at the final time point. Plots b and c were produced using the R package *ggmuller* [78]. Image a is reproduced from [77] under the terms of a Creative Commons Attribution License and with the approval of Paul Macklin. Plots b-d are reproduced from [13] under the terms of a Creative Commons Attribution License.

implementation. Diffusible factor concentrations can be shown outside the tumour using a colour gradient and inside the tumour by adjusting brightness [15]. We can apply the same method to off-lattice models by specifying a grid and assigning each grid square a value that summarises the states of all points within the square. Given multi-level spatial structure, we can represent the most abundant state in each deme [12].

Illustrating three-dimensional information is more technically demanding as we need to project the object onto a two-dimensional plane, determine the visible surface, and add shading (as in Figure 5a). Suitable computational methods include rasterisation and ray tracing, which can be performed in R and Python or using dedicated software, such as POV-Ray. Further details can be found in the PhysiCell documentation (see appendix). A much simpler solution is to plot only two-dimensional slices.

6.2 Visualising evolutionary dynamics

Muller plots represent subpopulation dynamics and phylogeny, while disregarding spatial information. The horizontal axis represents time and the vertical axis corresponds to subpopulation frequency. Each subpopulation is depicted as a shaded area emerging from its immediate ancestor (Figure 5b). Fish plots are similar but show population size rather than frequency (Figure 5c). Software packages for producing these plots include ggmuller [78] and EvoFreq [79].

6.3 Phenotype space plots

In a phenotype space plot, the axes correspond to continuous traits such as cell fitness, metabolic type, and degree of drug resistance, and each point represents a cell. We can visualize phenotypic evolution by animating phenotype space plots from a series of time points. Robertson-Tessi and colleagues pioneered the use of these plots in cancer research in their 2015 study of the effects of metabolic heterogeneity on tumour growth [48].

7 Reproducibility

Reproducibility is a cornerstone of the scientific method. A reproducible modelling study not only allows others to easily regenerate its results but also permits further data processing, downstream analysis of generated data, generation of summary statistics, ease of production for visual representations or plots, and even adaptation of the existing model for novel purposes.

7.1 Principles of reproducible research

Gundersen [80] describes three categories of reproducibility:

- Outcome reproducibility: The reproduction experiment's result matches the original. If the same analysis of the result is performed, the same conclusions can be drawn, and the original hypothesis is supported by both experiments.
- Analysis reproducibility: The reproduction experiment's result differs from the original, but if the same analysis method is used, the interpretation of the results still matches the original.
- Interpretation reproducibility: The reproduction experiment's outcomes and the analysis of said outcomes both differ, but the interpretation matches the original interpretation.

Computational modelling studies should typically aim for the highest standard of outcome reproducibility. If care is taken to construct a well-packaged computational study in a controlled digital environment, then in principle, given a suitable machine, the study should be easily reproduced exactly. This entails not only comprehensively explaining methods, results, analyses, and interpretation, but also sharing the model code and scripts used at every step of pre-processing and analysis, providing a detailed description of how to execute the code, and sharing any associated data and parameterisation and configuration files.

In their outline of best practices to observe throughout a computational research project, Sandve and colleagues [81] advocate tracking how every result is produced and reporting intermediate results as well as final outcomes. To make code easier to reproduce, one should catalogue the versions of software used at every point, record the seeds used in any random number generation, and implement version control [82]. Manual data manipulation should be avoided in favour of using automated methods to reformat and process raw data files. The raw data used to produce summary data plots should be easily accessible to facilitate easy plot reproduction and to allow readers to check individual data points. Textual descriptions of methods and results should link to the associated raw data and code so that a reader can easily follow all the steps leading to interpretations. Lastly, modellers are highly encouraged to share each full study, ideally with a dedicated public server. One such research-oriented database is zenodo [83], where scientists may freely upload their research output permanently as a citeable piece of software.

7.2 Workflow managers, package managers and containers

A complex computational model will often require multiple steps to be carried out in sequence. If a high-performance computing (HPC) cluster is required to run the model efficiently – as is typical for complex models – it is essential to utilise a workflow manager to properly orchestrate the steps [84]. Open-source workflow managers allow researchers to package a model into a reproducible, cross-platform workflow. Nextflow [85] and Snakemake [86] are among the most popular workflow managers with several published pipelines [87, 88, 89, 90], strong community support, and extensive documentation, giving users flexibility when designing their own custom pipelines. Snakemake is based on Python, a popular language among computational biologists and bioinformaticians. Nextflow uses the Java-based language Groovy, which has a Python-style structure and is relatively easy to for Python users to learn. Both also enable automatic parallelisation for HPC clusters, which can be essential for complex SABMs or for running multiple instances of smaller models simultaneously.

Another option is to utilise container technologies, considered by many to be the gold standard in computational research. These are less computationally demanding than running an application on a computer directly or using a virtual machine and so permit faster deployment, patching, and scaling. Containers also allow users to deploy the application on multiple operating systems or machines without reformatting, and will run the application the same way no matter where they are deployed [91]. Docker [92] is a popular container design platform which permits packaging applications into distribution-independent containers. Another option, Bioconda [93], enables easy dependency management, and can be deployed inside a container.

7.3 Extendable modelling platforms

For many research projects, the easiest option can be to build on an existing open-source agent-based modelling platform (see appendix for a brief guide). Some of these platforms – such as Chaste [94], CompuCell3D [95], HAL [51] and PhysiCell [77] – excel in simulating off-lattice cell populations in complex microenvironments. Others, such as demon [96] (which has an automated computational workflow, Warlock [14]), J-SPACE [97] and SMITH [98], focus on efficient modelling of evolutionary dynamics. Several are modular platforms, which facilitate reproducibility by making it easy to create and share extensions of the generic software. Nevertheless, even the most flexible platform is necessarily based on certain fundamental assumptions, structures, and algorithms. If we want to create an especially innovative model, requiring several novel components that pre-existing modelling platforms lack, then we might find it best to start from scratch. In principle, specialist rather than generalist models permit greater optimisation in terms of memory demands and execution time.

7.4 FAIR principles in data management

As the volume of publicly available research data has been growing exponentially in recent decades [99], proper digital data management and annotation is recognized as an essential step in computational research – crucial for research reproducibility. Most notably, the FAIR principles have become a cornerstone in modern data management, particularly in the realms of scientific and research data [100]. FAIR is an acronym that encapsulates a set of guiding principles: Findable, Accessible, Interoperable, and Reusable. To be FAIR, data must first be Findable, meaning that it is easy for both humans and machines to discover, thanks to comprehensive metadata and proper indexing. Data should be Accessible, ensuring that access rights and permissions are clear and well-defined, thus minimizing barriers to entry. Interoperable data is structured in a way that allows integration with other datasets by adhering to common standards, formats, and vocabularies. Lastly, data should be Reusable, with thorough documentation, contextual information, and availability in a format that facilitates easy replication and reuse. Altogether, the FAIR principles serve as a framework for enhancing data sharing, management, and collaboration, ultimately driving scientific progress and fostering open science initiatives. Major organisations that have embraced FAIR guidelines include the European Open Science Cloud [101], the European Life-Science Infrastructure for Biological Information [102], the US National Institutes of Health, and the Global Alliance for Genomics and Health [103].

Discussion

Having surveyed the numerous choices that arise in any SABM project, we are faced with a problem: how can we choose the most appropriate model? In tumour evolution research, unlike in much of physics and engineering, there is no standard approach. Rather we must tailor a model to each research question by considering which components, events and interactions must be included, how far each aspect can be parameterised with available data, and the limits of our computational resources. It is essential to build on a sound understanding of the biological system and of the questions that matter to biologists and clinicians. Ideally this knowledge should come through close collaboration with empirical researchers throughout the model development process.

A general principle is that model complexity should match the complexity only of the phenomena of interest. We need not employ an off-lattice hybrid SABM if a simple cellular automaton with only a few basic update rules can demonstrate the same principle. Attempting to represent every component of a biological system is not only computationally impractical but also risks overfitting and hinders explainability. Simpler models have many merits. They are easier to falsify and have fewer sources of potential error. They reduce researcher degrees of freedom and curb the tweaking of parameters to support a pet hypothesis. They are more mathematically tractable and easier to analyse. Perhaps most importantly, a simple model has wider applicability and can be more readily generalised, adapted or extended to answer new questions. More complicated models should be preferred only if the biological system is especially well understood or if simpler models have been tested and shown to be inadequate.

The greater difficulty – all too easily overlooked – is in choosing among a multitude of plausible simple models. An Eden growth model, for example, is arguably no more parsimonious than a spatial branching process or a spatial Moran process, which generate very different evolutionary dynamics. It is debatable whether the greater popularity of Eden growth models can be justified on biological grounds and is not simply due to them being easier to program.

Model design remains a challenge for even the most experienced researchers. One of the nine overarching themes in a recent review of key questions concerning the ecology and evolution of cancer [104] was that we do not yet know which mathematical and computational models are the most useful. In another recent survey of cancer adaptive therapy modelling [105], four of the eleven key open questions were related to identifying appropriate mathematical models. When it comes to SABMs, the main limitations are twofold. First, we typically lack sufficient data to design and parameterise SABMs of large tumours. Second, routinely simulating much more than a few million individual cells (corresponding to no more than half a cubic centimetre of tumour) is computationally impractical. To some extent, these problems have technological solutions. Multi-region sequencing, spatial multi-omics, digital pathology, and other modern methods are producing ever more detailed spatial tumour data. Accessible computing power continues to grow. But progress will also depend on developing smarter models.

Instead of drawing conclusions from a single SABM, we might do better to consider ensembles of models with diverse structures, algorithms, and underlying assumptions. Much as in hurricane forecasting [106], we can be more confident when many models converge on the same prediction. Another important direction is to develop coarse-grained models that can simulate tumour evolution as accurately as cell-level SABMs but with much greater computational efficiency. Rather than cell division, death, mutation and dispersal rates, coarse-grained models depend on macroscopic parameters such as the arrival rate of consequential clones, clonal expansion speeds, and large-scale microenvironmental heterogeneity. A potential way forward is to combine mathematical analysis of the relevant stochastic processes to determine appropriate approximations [107], and machine learning methods to infer the parameter values. SABMs capable of accurately simulating the evolution of entire tumours could have wide-ranging applications, not least in patient-specific clinical forecasting.

References

[1] Matias Casás-Selves and James DeGregori. How cancer shapes evolution and how evolution shapes cancer. *Evolution: Education and outreach*, 4:624–634, 2011.

[2] Lauren MF Merlo, John W Pepper, Brian J Reid, and Carlo C Maley. Cancer as an evolutionary and ecological process. *Nature reviews cancer*, 6(12):924–935, 2006.

[3] Jing Yang, Tong-Jun Zhao, Chang-Qing Yuan, Jing-Hui Xie, and Fang-Fang Hao. A nonlinear competitive model of the prostate tumor growth under intermittent androgen suppression. *Journal of theoretical biology*, 404:66–72, 2016.

[4] David Basanta, Jacob G Scott, Russ Rockne, Kristin R Swanson, and Alexander RA Anderson. The role of idh1 mutated tumour cells in secondary glioblastomas: an evolutionary game theoretical view. *Physical biology*, 8(1):015016, 2011.

[5] Kaveh Danesh, Rick Durrett, Laura J Havrilesky, and Evan Myers. A branching process model of ovarian cancer. *Journal of theoretical biology*, 314:10–15, 2012.

[6] Rick Durrett, Jasmine Foo, Kevin Leder, John Mayberry, and Franziska Michor. Evolutionary dynamics of tumor progression with random fitness values. *Theoretical population biology*, 78(1):54–66, 2010.

[7] Jeffrey West, Zaki Hasnain, Paul Macklin, and Paul K Newton. An evolutionary model of tumor cell kinetics and the emergence of molecular heterogeneity driving gompertzian growth. *SIAM review*, 58(4):716–736, 2016.

[8] Richard Durrett, Jasmine Foo, and Kevin Leder. Spatial moran models, ii: cancer initiation in spatially structured tissue. *Journal of mathematical biology*, 72:1369–1400, 2016.

[9] Serhii Aif, Nico Appold, Lucas Kampman, Oskar Hallatschek, and Jona Kayser. Evolutionary rescue of resistant mutants is governed by a balance between radial expansion and selection in compact populations. *Nature Communications*, 13(1):7916, 2022.

[10] Daniel K Saha, Alexander RA Anderson, Luis Cisneros, and Carlo C Maley. In silico investigations of adaptive therapy using a single cytotoxic or a single cytostatic drug. *bioRxiv*, pages 2023–05, 2023.

[11] Joshua A Bull and Helen M Byrne. Quantification of spatial and phenotypic heterogeneity in an agent-based model of tumour-macrophage interactions. *PLOS Computational Biology*, 19(3):e1010994, 2023.

[12] Robert Noble, Dominik Burri, Cécile Le Sueur, Jeanne Lemant, Yannick Viossat, Jakob Nikolas Kather, and Niko Beerenwinkel. Spatial structure governs the mode of tumour evolution. *Nature ecology & evolution*, 6(2):207–217, 2022.

[13] Robert Noble, John T Burley, Cécile Le Sueur, and Michael E Hochberg. When, why and how tumour clonal diversity predicts survival. *Evolutionary applications*, 13(7):1558–1568, 2020.

[14] Maciej Bak, Blair Colyer, Veselin Manojlović, and Robert Noble. Warlock: an automated computational workflow for simulating spatially structured tumour evolution. *arXiv preprint arXiv:2301.07808*, 2023.

[15] Katarina Bacevic, Robert Noble, Ahmed Soffar, Orchid Wael Ammar, Benjamin Boszonyik, Susana Prieto, Charles Vincent, Michael E Hochberg, Liliana Krasinska, and Daniel Fisher. Spatial competition constrains resistance to targeted cancer therapy. *Nature communications*, 8(1):1995, 2017.

[16] Maximilian AR Strobl, Jill Gallaher, Jeffrey West, Mark Robertson-Tessi, Philip K Maini, and Alexander RA Anderson. Spatial structure impacts adaptive therapy by shaping intra-tumoral competition. *Communications medicine*, 2(1):46, 2022.

[17] Joel L Schiff. *Cellular automata: a discrete view of the world*. John Wiley & Sons, 2011.

[18] Pierre-Yves Louis and Francesca R Nardi. Probabilistic cellular automata. *Emergence, Complexity, Computation*, 27, 2018.

[19] Murray Eden. A two-dimensional growth process. *Dynamics of fractal surfaces*, 4:223–239, 1961.

[20] R Jullien and R Botet. Scaling properties of the surface of the eden model in d= 2, 3, 4. *Journal of Physics A: Mathematical and general*, 18(12):2279, 1985.

[21] Haider Tari, Ketty Kessler, Nick Trahearn, Benjamin Werner, Maria Vinci, Chris Jones, and Andrea Sottoriva. Quantification of spatial subclonal interactions enhancing the invasive phenotype of pediatric glioma. *Cell Reports*, 40(9), 2022.

[22] Sara Hamis, James Yates, Mark AJ Chaplain, and Gibin G Powathil. Targeting cellular dna damage responses in cancer: an in vitro-calibrated agent-based model simulating monolayer and spheroid treatment responses to atr-inhibiting drugs. *Bulletin of Mathematical Biology*, 83:1–21, 2021.

[23] Bartlomiej Waclaw, Ivana Bozic, Meredith E Pittman, Ralph H Hruban, Bert Vogelstein, and Martin A Nowak. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525(7568):261–264, 2015.

[24] Maya A. Lewinsohn, Trevor Bedford, Nicola F. Müller, and Alison F. Feder. State-dependent evolutionary models reveal modes of solid tumour growth. *Nature Ecology & Evolution*, 7(4):581–596, 2023.

[25] Xiao Fu, Yue Zhao, Jose I. Lopez, Andrew Rowan, Lewis Au, Annika Fendler, Steve Hazell, Hang Xu, Stuart Horswell, Scott T. C. Shepherd, Charlotte E. Spencer, Lavinia Spain, Fiona Byrne, Gordon Stamp, Tim O'Brien, David Nicol, Marcellus Augustine, Ashish Chandra, Sarah Rudman, Antonia Toncheva, Andrew J. S. Furness, Lisa Pickering, Santosh Kumar, Dow-Mu Koh, Christina Messiou, Derfel ap Dafydd, Matthew R. Orton, Simon J. Doran, James Larkin, Charles Swanton, Erik Sahai, Kevin Litchfield, Samra Turajlic, Ben Challacombe, Simon Chowdhury, William Drake, Archana Fernando, Nicos Fotiadis, Emine Hatipoglu, Karen Harrison-Phipps, Peter Hill, Catherine Horsfield, Teresa Marafioti, Jonathon Olsburgh, Alexander Polson, Sergio Quezada, Mary Varia, Hema Verma, Paul A. Bates, and on behalf of the TRACERx Renal Consortium. Spatial patterns of tumour growth impact clonal diversification in a computational model and the tracerx renal study. *Nature Ecology & Evolution*, 6(1):88–102, 2022.

[26] Nick Jagiella, Benedikt Müller, Margareta Müller, Irene E Vignon-Clementel, and Dirk Drasdo. Inferring growth control mechanisms in growing multi-cellular spheroids of nsclc cells from spatial-temporal image data. *PLoS computational biology*, 12(2):e1004412, 2016.

[27] Ketevan Chkhaidze, Timon Heide, Benjamin Werner, Marc J Williams, Weini Huang, Giulio Caravagna, Trevor A Graham, and Andrea Sottoriva. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *PLoS computational biology*, 15(7):e1007243, 2019.

[28] Richard Durrett. Random graph dynamics, 2007.

[29] Trevor Williams and Rolf Bjerknes. Stochastic model for abnormal clone spread through epithelial basal layer. *Nature*, 236(5340):19–21, 1972.

[30] François Graner and James A Glazier. Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical review letters*, 69(13):2013, 1992.

[31] Nicholas J Savill and Paulien Hogeweg. Modelling morphogenesis: from single cells to crawling slugs. *Journal of theoretical biology*, 184(3):229–235, 1997.

[32] András Szabó and Roeland MH Merks. Cellular potts modeling of tumor growth, tumor invasion, and tumor evolution. *Frontiers in oncology*, 3:87, 2013.

[33] Jonathan F Li and John Lowengrub. The effects of cell compressibility, motility and contact inhibition on the growth of tumor cell clusters using the cellular potts model. *Journal of theoretical biology*, 343:79–91, 2014.

[34] Andreas Deutsch and Sabine Dormann. *Mathematical modeling of biological pattern formation*. Springer, 2005.

[35] Andreas Deutsch, Josué Manik Nava-Sedeño, Simon Syga, and Haralampos Hatzikirou. Bio-lgca: a cellular automaton modelling class for analysing collective cell migration. *PLoS computational biology*, 17(6):e1009066, 2021.

[36] Pahini Pandya, Jose L Orgaz, and Victoria Sanz-Moreno. Modes of invasion during tumour dissemination. *Molecular oncology*, 11(1):5–27, 2017.

[37] Alessandro Lugli, Inti Zlobec, Martin D Berger, Richard Kirsch, and Iris D Nagtegaal. Tumour budding in solid cancers. *Nature Reviews Clinical Oncology*, 18(2):101–115, 2021.

[38] Sean L Preston, Wai-Man Wong, Annie On-On Chan, Richard Poulsom, Rosemary Jeffery, Robert A Goodlad, Nikki Mandir, George Elia, Marco Novelli, Walter F Bodmer, et al. Bottom-up histogenesis of colorectal adenomas: origin in the monocryptal adenoma and initial expansion by crypt fission. *Cancer research*, 63(13):3819–3825, 2003.

[39] Patrick Alfred Pierce Moran. Random processes in genetics, 1958.

[40] Andrea Sottoriva, Haeyoun Kang, Zhicheng Ma, Trevor A Graham, Matthew P Salomon, Junsong Zhao, Paul Marjoram, Kimberly Siegmund, Michael F Press, Darryl Shibata, et al. A big bang model of human colorectal tumor growth. *Nature genetics*, 47(3):209–216, 2015.

[41] Kimberly D Siegmund, Paul Marjoram, Yen-Jung Woo, Simon Tavaré, and Darryl Shibata. Inferring clonal expansion and cancer stem cell dynamics from dna methylation patterns in colorectal cancers. *Proceedings of the National Academy of Sciences*, 106(12):4828–4833, 2009.

[42] Niko Beerenwinkel, Roland F Schwarz, Moritz Gerstung, and Florian Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25, 2015.

[43] Benjamin Franz, Chuan Xue, Kevin J Painter, and Radek Erban. Travelling waves in hybrid chemotaxis models. *Bulletin of mathematical biology*, 76:377–400, 2014.

[44] Paul Macklin, Mary E Edgerton, Alastair M Thompson, and Vittorio Cristini. Patient-calibrated agent-based modelling of ductal carcinoma in situ (dcis): from microscopic measurements to macroscopic predictions of clinical progression. *Journal of theoretical biology*, 301:122–140, 2012.

[45] Philip Gerlee and Alexander RA Anderson. Modelling evolutionary cell behaviour using neural networks: application to tumour growth. *Biosystems*, 95(2):166–174, 2009.

[46] Jill A Gallaher, Pedro M Enriquez-Navas, Kimberly A Luddy, Robert A Gatenby, and Alexander RA Anderson. Spatial heterogeneity and evolutionary dynamics modulate time to recurrence in continuous and adaptive cancer therapies. *Cancer research*, 78(8):2127–2139, 2018.

[47] Keith S Hoek, Ossia M Eichhoff, Natalie C Schlegel, Udo Döbbeling, Nikita Kobert, Leo Schaerer, Silvio Hemmi, and Reinhard Dummer. In vivo switching of human melanoma cells between proliferative and invasive states. *Cancer research*, 68(3):650–656, 2008.

[48] Mark Robertson-Tessi, Robert J Gillies, Robert A Gatenby, and Alexander RA Anderson. Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. *Cancer research*, 75(8):1567–1579, 2015.

[49] Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

[50] Ryan O Schenck, Eunjung Kim, Rafael R Bravo, Jeffrey West, Simon Leedham, Darryl Shibata, and Alexander RA Anderson. Homeostasis limits keratinocyte evolution. *Proceedings of the National Academy of Sciences*, 119(35):e2006487119, 2022.

[51] Jeffrey West. Hybrid automata library. https://halloworld.org.

[52] Robert Noble. An eden model in r (with optional mutation of division rates). https://github.com/robjohnnoble/EdenModel, 2018.

[53] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[54] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of chemical physics*, 115(4):1716–1733, 2001.

[55] Takayuki Osogami and Mor Harchol-Balter. Closed form solutions for mapping general distributions to quasi-minimal ph distributions. *Performance Evaluation*, 63(6):524–552, 2006.

[56] Giulia Belluccini, Martín López-García, Grant Lythe, and Carmen Molina-París. Counting generations in birth and death processes with competing erlang and exponential waiting times. *Scientific Reports*, 12(1):11289, 2022.

[57] Nicole M Anderson and M Celeste Simon. The tumor microenvironment. *Current Biology*, 30(16):R921–R925, 2020.

[58] MA Chaplain and AR Anderson. Mathematical modelling, simulation and prediction of tumour-induced angiogenesis. *Invasion & metastasis*, 16(4-5):222–234, 1996.

[59] An-Shen Qi, Xiang Zheng, Chan-Ying Du, and Bao-Sheng An. A cellular automaton model of cancerous growth. *Journal of theoretical biology*, 161(1):1–12, 1993.

[60] Aalpen A Patel, Edward T Gawlinski, Susan K Lemieux, and Robert A Gatenby. A cellular automaton model of early tumor growth and invasion: the effects of native tissue vascularity and increased anaerobic tumor metabolism. *Journal of theoretical biology*, 213(3):315–331, 2001.

[61] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[62] Ahmadreza Ghaffarizadeh, Samuel H Friedman, and Paul Macklin. Biofvm: an efficient, parallelized diffusive transport solver for 3-d biological simulations. *Bioinformatics*, 32(8):1256–1258, 2016.

[63] Rafael R Bravo, Etienne Baratchart, Jeffrey West, Ryan O Schenck, Anna K Miller, Jill Gallaher, Chandler D Gatenbee, David Basanta, Mark Robertson-Tessi, and Alexander RA Anderson. Hybrid automata library: A flexible platform for hybrid modeling with real-time visualization. *PLoS computational biology*, 16(3):e1007635, 2020.

[64] Joseph J Casciari, Stratis V Sotirchos, and Robert M Sutherland. Mathematical modelling of microenvironment and growth in emt6/ro multicellular tumour spheroids. *Cell proliferation*, 25(1):1–22, 1992.

[65] Yangjin Kim, Magdalena A Stolarska, and Hans G Othmer. A hybrid model for tumor spheroid growth in vitro i: theoretical development and early results. *Mathematical Models and Methods in Applied Sciences*, 17(supp01):1773–1798, 2007.

[66] David Robert Grimes, Catherine Kelly, Katarzyna Bloch, and Mike Partridge. A method for estimating the oxygen consumption rate in multicellular tumour spheroids. *Journal of The Royal Society Interface*, 11(92):20131124, 2014.

[67] Joseph J Casciari, Stratis V Sotirchos, and Robert M Sutherland. Variations in tumor cell growth rates and metabolism with oxygen concentration, glucose concentration, and extracellular ph. *Journal of cellular physiology*, 151(2):386–394, 1992.

[68] David Robert Grimes, Alexander G Fletcher, and Mike Partridge. Oxygen consumption dynamics in steady-state tumour models. *Royal Society open science*, 1(1):140080, 2014.

[69] Arturo Araujo and David Basanta. Hybrid discrete-continuum cellular automaton (hca) model of prostate to bone metastasis. *bioRxiv*, page 043620, 2016.

[70] Arturo Araujo, Leah M Cook, Conor C Lynch, and David Basanta. Size matters: Metastatic cluster size and stromal recruitment in the establishment of successful prostate cancer to bone metastases. *Bulletin of mathematical biology*, 80:1046–1058, 2018.

[71] Arturo Araujo, Leah M Cook, Conor C Lynch, and David Basanta. An integrated computational model of the bone microenvironment in bone-metastatic prostate cancer. *Cancer research*, 74(9):2391–2401, 2014.

[72] Andrea Sottoriva and Simon Tavaré. Integrating approximate bayesian computation with complex agent-based models for cancer research. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 57–66. Springer, 2010.

[73] Colin G Cess and Stacey D Finley. Calibrating agent-based models to tumor images using representation learning. *PLOS Computational Biology*, 19(4):e1011070, 2023.

[74] Il'ya Meerovich Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4):784–802, 1967.

[75] Michael D McKay, Richard J Beckman, and William J Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.

[76] Sara Hamis, Stanislav Stratiev, and Gibin G Powathil. Uncertainty and sensitivity analyses methods for agent-based mathematical models: An introductory review. *The Physics of Cancer: Research Advances*, pages 1–37, 2021.

[77] Ahmadreza Ghaffarizadeh, Randy Heiland, Samuel H Friedman, Shannon M Mumenthaler, and Paul Macklin. Physicell: An open source physics-based cell simulator for 3-d multicellular systems. *PLoS computational biology*, 14(2):e1005991, 2018.

[78] Robert Noble. ggmuller: Create muller plots of evolutionary dynamics. r package version 0.5.3. https://CRAN.R-project.org/package=ggmuller, 2019.

[79] Chandler D Gatenbee, Ryan O Schenck, Rafael R Bravo, and Alexander RA Anderson. Evofreq: visualization of the evolutionary frequencies of sequence and model data. *BMC bioinformatics*, 20:1–4, 2019.

[80] Odd Erik Gundersen. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A*, 379(2197):20200210, 2021.

[81] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS computational biology*, 9(10):e1003285, 2013.

[82] Michael A Heroux and James M Willenbring. Barely sufficient software engineering: 10 practices to improve your cse software. In *2009 ICSE workshop on software engineering for computational science and engineering*, pages 15–21. IEEE, 2009.

[83] Zenodo. https://zenodo.org/, 2023.

[84] Laura Wratten, Andreas Wilm, and Jonathan Göke. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature methods*, 18(10):1161–1168, 2021.

[85] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, et al. Sustainable data analysis with snakemake. *F1000Research*, 10, 2021.

[86] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.

[87] Silas Kieser, Joseph Brown, Evgeny M Zdobnov, Mirko Trajkovski, and Lee Ann McCue. Atlas: a snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC bioinformatics*, 21:1–8, 2020.

[88] Martin Hölzer and Manja Marz. Poseidon: a nextflow pipeline for the detection of evolutionary recombination events and positive selection. *Bioinformatics*, 37(7):1018–1020, 2021.

[89] Qi Zhao, Yu Sun, Dawei Wang, Hongwan Zhang, Kai Yu, Jian Zheng, and Zhixiang Zuo. Lncpipe: A nextflow-based pipeline for identification and analysis of long non-coding rnas from rna-seq data. *J Genet Genomics*, 45(7):399–401, 2018.

[90] MacIntosh Cornwell, Mahesh Vangala, Len Taing, Zachary Herbert, Johannes Köster, Bo Li, Hanfei Sun, Taiwen Li, Jian Zhang, Xintao Qiu, et al. Viper: Visualization pipeline for rna-seq, a snakemake workflow for efficient and complete rna-seq analysis. *BMC bioinformatics*, 19:1–14, 2018.

[91] David Moreau, Kristina Wiebels, and Carl Boettiger. Containers for computational reproducibility. *Nature Reviews Methods Primers*, 3(1):50, 2023.

[92] Dirk Merkel et al. Docker: lightweight linux containers for consistent development and deployment. *Linux j*, 239(2):2, 2014.

[93] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A Chapman, Jillian Rowe, Christopher H Tomkins-Tinch, Renan Valieris, Johannes Köster, and Bioconda Team. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, 15(7):475–476, 2018.

[94] Gary R Mirams, Christopher J Arthurs, Miguel O Bernabeu, Rafel Bordas, Jonathan Cooper, Alberto Corrias, Yohan Davit, Sara-Jane Dunn, Alexander G Fletcher, Daniel G Harvey, et al. Chaste: an open source c++ library for computational physiology and biology. *PLoS computational biology*, 9(3):e1002970, 2013.

[95] Maciej H Swat, Gilberto L Thomas, Julio M Belmonte, Abbas Shirinifard, Dimitrij Hmeljak, and James A Glazier. Multi-scale modeling of tissues using compucell3d. In *Methods in cell biology*, volume 110, pages 325–366. Elsevier, 2012.

[96] Robert Noble. demon. `https://github.com/robjohnnoble/demon_model`, 2019.

[97] Fabrizio Angaroni, Alessandro Guidi, Gianluca Ascolani, Alberto d'Onofrio, Marco Antoniotti, and Alex Graudenzi. J-space: a julia package for the simulation of spatial models of cancer evolution and of sequencing experiments. *BMC bioinformatics*, 23(1):269, 2022.

[98] Adam Streck, Tom L Kaufmann, and Roland F Schwarz. Smith: spatially constrained stochastic model for simulation of intra-tumour heterogeneity. *Bioinformatics*, 39(3):btad102, 2023.

[99] Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/, August 2023.

[100] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[101] The european open science cloud (eosc). https://eosc-portal.eu/about/eosc.

[102] European life-science infrastructure for biological information. https://elixir-europe.org.

[103] Global alliance for genomics and health. https://www.ga4gh.org.

[104] Antoine M Dujon, Athena Aktipis, Catherine Alix-Panabières, Sarah R Amend, Amy M Boddy, Joel S Brown, Jean-Pascal Capp, James DeGregori, Paul Ewald, Robert Gatenby, et al. Identifying key questions in the ecology and evolution of cancer. *Evolutionary Applications*, 14(4):877–892, 2021.

[105] Jeffrey West, Fred Adler, Jill Gallaher, Maximilian Strobl, Renee Brady-Nicholls, Joel Brown, Mark Roberson-Tessi, Eunjung Kim, Robert Noble, Yannick Viossat, et al. A survey of open questions in adaptive therapy: Bridging mathematics and clinical translation. *Elife*, 12:e84263, 2023.

[106] Thomas M Hamill, Michael J Brennan, Barbara Brown, Mark DeMaria, Edward N Rappaport, and Zoltan Toth. Noaa's future ensemble-based hurricane forecast products. *Bulletin of the American Meteorological Society*, 93(2):209–220, 2012.

[107] Alexander Stein, Ramanarayanan Kizhuttil, Maciej Bak, and Robert Noble. The ubiquity of clonal interference in cancer and other range expansions. *bioRxiv*, 2024.

Data archiving statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.