LETTERS IN BIOMATHEMATICS
An International Journal

Intercollegiate Biomathematics Alliance

RESEARCH ARTICLE

OPEN ACCESS

Optimization of Agent-Based Models Through Coarse-Graining: A Case Study in Microbial Ecology

Sherli Koshy-Chenthittayil[a], Pedro Mendes[b], Reinhard Laubenbacher[c]

[a]Center for Quantitative Medicine, University of Connecticut Health Center (koshychenthittayil@uchc.edu); [b]Center for Quantitative Medicine and Center for Cell Analysis and Modeling, University of Connecticut Health Center (pmendes@uchc.edu); [c]Department of Medicine, University of Florida (reinhard.laubenbacher@medicine.ufl.edu)

ABSTRACT

Optimization and control are important objectives across biology and biomedicine, and mathematical models are a key enabling technology. This paper reports a computational study of model-based multi-objective optimization in the setting of microbial ecology, using agent-based models. This modeling framework is well-suited to the field, but is not amenable to standard control-theoretic approaches. Furthermore, due to computational complexity, simulation-based optimization approaches are often challenging to implement. This paper presents the results of an approach that combines control-dependent coarse-graining with Pareto optimization, applied to two models of multi-species bacterial biofilms. It shows that this approach can be successful for models whose computational complexity prevents effective simulation-based optimization.

ARTICLE HISTORY
Received November 9, 2020
Accepted August 20, 2021

KEYWORDS
optimization,
agent-based models,
Pareto frontier,
microbial biofilms

1 Introduction

Agent-based models (ABMs) provide an increasingly popular modeling platform across fields, such as ecology, social sciences, and the life sciences. Their use has grown significantly to model different processes in biomedicine and human health (An et al., 2009; Giabbanelli and Crutzen, 2017; Nealon and Moreno, 2003), as they provide an intuitive rule-based design that allows for the explicit representation of individual entities, such as cells in the human body or individual members of a community, and the inclusion of highly heterogeneous spatial environments. Global model dynamics emerge from the totality of many different individual interactions between agents and those of agents with the local spatial environment. A wide range of examples include models of microbial communities (Lardon et al., 2011; Li et al., 2019), the heroin market (Heard et al., 2014), traffic and pollution (Forhead and Huynh, 2018), and aspects of climate change (Tinker and Wilensky, 2007). A significant drawback of ABMs is that they are, in their essence, computational algorithms that are not equation-based, which makes it difficult to analyze their dynamics and other properties.

Many problems in biomedicine are focused on control. For example, dietary control can lead to a gut microbiome better optimized for health, possibly avoiding irritable bowel syndrome, cardiovascular conditions, and other diseases (Shreiner et al., 2015). Another example is in controlling the growth of tumor cells in cancer patients using immunotherapy (Couzin-Frankel, 2013). Mathematical and computational models represent a key enabling technology to solve such problems, especially in settings where even trial-and-error approaches are not feasible. However, in contrast to the case of equation-based models, such as systems of ordinary differential equations, few general mathematical techniques are available for ABMs, beyond extensive model simulation. For complex models, even that becomes challenging as the computational cost of simulating a model can be quite high (An et al., 2009). For example, over two weeks of computation time were required for the optimization of the multi-heterotroph model used in the present study. Several different approaches have been developed to circumvent this problem, ranging from various heuristic approaches to approximation of agent-based models by other model types. A general discussion of this topic can be found in (An et al., 2017). Only heuristic approaches are available at this time to solve this quite difficult problem (see below for a brief survey), and a more in-depth study of their applicability is needed to extract general principles.

In this paper we present a case study of an optimization approach discussed in (An et al., 2017), first proposed in (Oremland and Laubenbacher, 2015). It was demonstrated there with a more limited case study and much simpler models. It applies to an

CONTACT R. Laubenbacher reinhard.laubenbacher@medicine.ufl.edu

ABM, together with an optimization problem, taken here to be multi-objective, quite typical in many applications. The first step is a coarse-graining process resulting in a model more computationally tractable. The extent of coarse-graining is designed to preserve the nature of the control problem, in a well-defined manner. The control problem can then be solved for the coarse-grained model and the solution lifted back to the original model. A commonly used optimization technique for this purpose is Pareto optimization (Moore, 1897). It determines a set of solutions (Pareto frontier) that cannot be improved upon in terms of one objective without a sacrifice in at least one other objective. However, in principle, any other optimization technique could also be used for this purpose.

Figure 1 captures the basic idea of this approach. As in many situations, there are likely no "free lunches," in the sense that coarse-graining will reduce computational expense, but might also lose some of the optimality of solutions obtained from the coarse-grained model, when lifted to the original model. So the approach presented here is, by itself, a multi-objective optimization problem to reduce computational expense while maintaining fidelity of optimization.

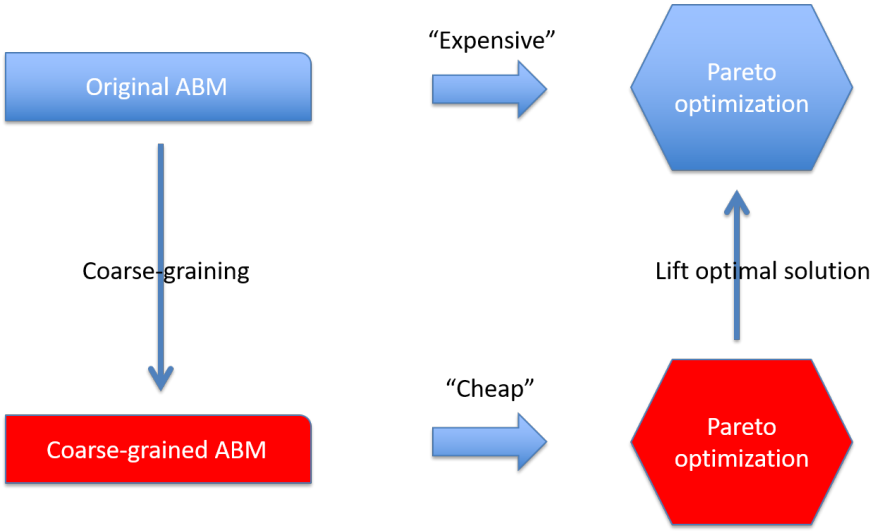

Figure 1: The approach used in this experimental study is as follows. Pareto optimization proceeds by simulating the model with many different settings. For the full model, this has typically a large computational cost. The approach we have implemented carries out coarse-graining first, then Pareto optimization with lower computational cost. Optimal solutions can then be lifted to the original ABM.

The biological focus of this study is on ABMs of heterogeneous biofilms. A biofilm forms when free floating planktonic microbial cells adhere to a damp surface. The cells begin excreting an extracellular polymeric substance (EPS) and start forming a matrix anchored on the surface. The bacterial population inside the matrix grows by consuming nutrients supplied by the surrounding environment (Wang and Zhang, 2010). Microbial biofilms can cause serious health problems in a variety of settings, from mucosal infections in immuno-compromised patients to medical implants and contamination of surfaces in hospitals. Therapeutic approaches are limited in most of these cases since the microbial cells embedded in the EPS matrix of the biofilm are largely protected from the action of antibiotics (Bryers, 2008). On the other hand, bacterial biofilms can also be beneficial, such as in wastewater treatment (Lazarova and Manem, 1995; Andersson, 2009). The fixed nature of the biofilms makes the microbes less susceptible than planktonic cultures to changes in environmental conditions within a wastewater treatment bioreactor.

In this paper, we carry out a more extensive computational study of the optimization method in (Oremland and Laubenbacher, 2015), using two different ABM models, comparing and contrasting the results for each. The ABM models chosen were of a harmful and a beneficial biofilm. The role of the biofilm dictated the choice of which outputs we wished to optimize. For example, if the biofilm was beneficial, we would like to maximize its thickness and cell numbers and if it is harmful, we would like to minimize both those measures. We first provide a brief overview of existing methods, in order to put our proposed method into the context of the literature on the subject. Next we describe the models we are focusing our investigation on, and a detailed description of the study design. Following are the results of the study and a discussion of their implications for the solution of optimization problems using ABMs.

2 Background

ABMs are typically stochastic algorithms that can be computationally complex, with large numbers of entities (agents) that interact based on a set of rules for each agent, and a typically heterogeneous spatial environment (Li et al., 2019; Bauer et al., 2017). They are notoriously difficult to analyze systematically. This includes problems like the determination of the optimal number of solutions to obtain stable variable trajectories without excessive computation time, or methods for sensitivity analysis on the parameter space. Each of Lee et al. (2015); Rhodes et al. (2016); Oyebamiji et al. (2017) provide some techniques to address these issues, based on coefficient of variation, effect size, multivariate stability, standardized regression coefficients, meta-modeling, variance-based decomposition, and the Sobol global sensitivity method.

One popular technique of optimizing ABMs is the use of statistical emulators. Emulators are statistical approximations of ABMs with regards to certain inputs and outputs of interest (Heard et al., 2015). The emulators can be Gaussian processes, which have been applied to study the Semi-Artificial Model of Population (Bijak et al., 2013), based on the "Wedding Ring" ABM of marriage formation, ABMs of microbial communities in water, and biofilms (Oyebamiji et al., 2017).

Another method is approximate Bayesian computation (ABC), which estimates the posterior distribution of a parameter of interest using observed data from multiple simulations (Heard et al., 2015). Oyebamiji et al. (2019) used a mixture of Gaussian processes and dynamic linear modeling to study the Newcastle University Frontiers in Engineering Biology (NUFEB) model.

Coarse-graining of ABMs has also been studied. In (Heard et al., 2014), a local heroin market agent-based model (IDMS-illicit drug market simulation) is reduced to increase efficiency and preserve the original model's statistical characteristics. The authors looked for outcomes that have an effect on the heroin market dynamics and the overall simulation. They then computed statistical approximations of the outcomes by fitting the data to a distribution or regression models. To compare the models, they ran the full and reduced models and produced trajectories of certain outcomes, and evaluated the goodness of fit between the full and coarse-grained models based on these outcomes.

An instance of model reduction can be found in (Willem et al., 2015). The authors compared two published ABMs for pandemic influenza: FluTE (Chao et al., 2010) and FRED (Grefenstette et al., 2013). They investigate data management, algorithmic procedures and parallelization of the models, carrying out reductions of the original model based on its structure using specific disease transmission models.

3 Methods

In this paper, we employ a multi-objective evolutionary algorithm (MOEA) to evaluate the Pareto frontier and Cohen's kappa value (Cohen, 1960, 1968) to determine if a particular coarse-graining and/or initial seed reduction preserves the key features of the optimization problem under consideration. This section describes each of these techniques and also gives a summary of the models investigated.

3.1 Agent-based simulation of microbial biofilms

The microbial biofilm models (see below) were simulated using the open source software iDynoMiCS (Lardon et al., 2011) from a Python interface. All simulations are controlled by evolutionary algorithms, using an interface between R and Python. All computations were run on a Windows 10 PC with a quad-core 64 bit Intel Xenon CPU with 32GB RAM. The versions of software used here were R 3.4.3, Python 3.7.0, and iDynoMiCS 1.2. All the relevant R code and iDynoMiCS protocol files are available at `https://github.com/skoshyc/optimization_Pareto.git`.

When we use the phrase "coarse-graining" in the remainder of the paper, we mean that we have reduced the number of spatial grid points of the model and/or its initial seed. Also, the term "reduced model" in the figures and tables will be a coarse-graining and/or an initial seed reduction of the original model.

3.2 Model coarse-graining and comparison

Before describing the coarse-graining algorithm, we define Cohen's Kappa coefficient. It was initially intended to measure the degree of agreement in different rankings of a set of objects by judges. It is given by

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \qquad \text{or equivalently} \qquad \kappa = \frac{f_0 - f_c}{N - f_c}$$

where p_0 is the proportion of rankings on which the judges agreed, p_c is the proportion of rankings for which agreement is expected by chance, f_0, f_c are the frequencies, and N is the total number of items (Cohen, 1960). The maximum value of κ is 1 which indicates perfect agreement between the judges. That is, the higher the value of κ the better the agreement among judges.

Cohen's weighted kappa κ_w is used when different kinds of disagreement are to be differentially weighted in the agreement index (Cohen, 1968). It is given by

$$\kappa_w = 1 - \frac{\sum v_{ij} p_{0ij}}{\sum v_{ij} p_{cij}}$$

where i, j range over the k categories which are being ranked, v_{ij} are the weights assigned to each entry of the $k \times k$ table, and p_0, p_c are the same as for Cohen's kappa. The weights can be determined based on the context but have to be assigned prior to the collection of the data.

Here we calculate Cohen's weighted kappa value in the R language using the function `kapp2` from the `irr` package with L2 distance metrics (option "squares" in that function) (Gamer et al., 2019). To determine if a particular reduction should be considered, we first calculate Cohen's kappa value to determine the level of "agreement" between the original and the reduced model. Using Latin hypercube sampling (R package `lhs` (Carnell, 2020)), combinations of the inputs were generated. Simulations for each of these combinations were run for both the original and reduced models and a particular output of interest was extracted, in this case the biofilm thickness. The Cohen's weighted kappa is then calculated between the ordinal list of the chosen output of the original and coarse-grained model. During these simulations, we also keep note of other outputs, like the species count and diversity index. We then go on to calculate the Pareto frontier using an evolutionary algorithm.

3.3 Pareto frontier

In multi-objective optimization, a Pareto frontier is a set of solutions for which improving one objective will be detrimental to at least one other objective (Van Veldhuizen and Lamont, 1998). Examples of a Pareto front using the `mtcars` dataset from R and the R package rPref (Roocks, 2016) is shown in Figure 2. In both the subfigures, 1 stands for points on the Pareto front and 0 stands for points which are not and the blue line connecting the points is the Pareto front for the particular optimization. The objectives in this example were to optimize the two variables *mpg* and *hp*. In Figure 2(a), the goal was to maximize the two objectives and that is why the Pareto front is in the upper half of the data. Similarly, for the minimization objective in Figure 2(b), the Pareto front is in the lower half.

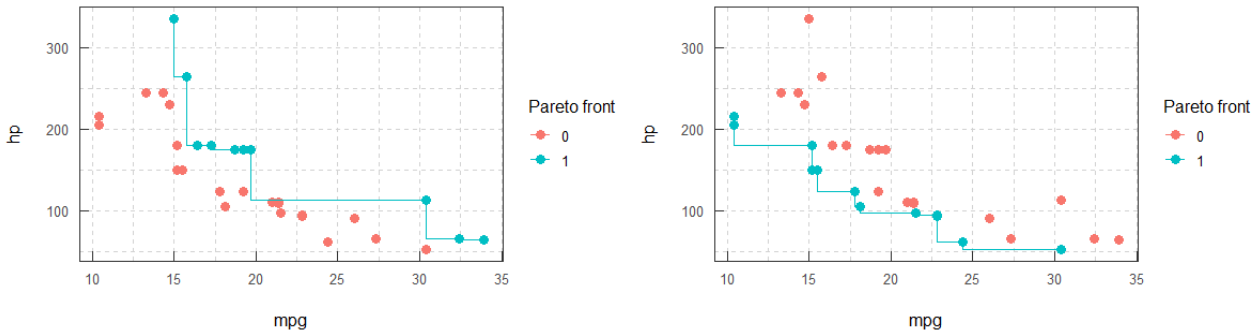

Figure 2: (a) An example of a Pareto front when the goal is to maximize two objectives. 1 stands for all points on the Pareto front and 0 stands for the points which are not. (b) Pareto front when the goal is to minimize the objectives.

3.4 Estimating the Pareto frontier

Evolutionary algorithms are popular multi-objective optimization tools. Some of the prominent methods are IBEA (Indicator-Based Evolutionary Algorithm), SPEA2 (Strength Pareto Evolutionary Algorithm 2), NSGA2 (Non-dominated Sorting Genetic Algorithm) (Small et al., 2011). The pseudocode in Algorithm 1 is based on the algorithm given in (Oremland and Laubenbacher, 2015) and the NSGA2 original algorithm (Deb et al., 2002). The code was written in R using the package `nsga2R` with modifications to incorporate the iDynoMiCS software.

A brief description of the pseudocode in Algorithm 1 is as follows. The inputs are solute concentrations and the outputs may be biofilm thickness, diversity index or the cell count of a particular microbal species. The ordinal list which was used to calculate the Cohen's kappa is taken as the initial set of solutions. A parent population of size 20 is chosen randomly from this set with a seed of 1234. (Note that beyond the choice of first parent population, mutations are chosen with random seeds.) The parent population is then ranked using the non-dominated sorting in the NSGA2 (Deb et al., 2002). The children population is generated by mutating the inputs. The xml file is updated by changing the mutated inputs and the simulation is run 10 times

Algorithm 1 Pseudocode for Pareto optimization

```
n = maximum number of generations
pop_size= size of population (say 20)
m = number of repetitions.
generate initial population of solutions                          ▷ store input
while gen < n do
    Use NSGA2 (Non-dominated Sorting Genetic Algorithm) to determine Pareto frontier (current_frontier) from current generation (current_pop)
    Generate child_pop by randomly mutating solute inputs of current_pop.
    Create new input xml file for each mutated solute concentrations.
    Run code with new input xml file for m repetitions.
    Take average of outputs across repetitions.
    Concatenate current_pop and child_pop.
    Use NSGA2 again to sort the concatenated list.
    new_pop= first pop_size entries of sorted list, new_frontier = Pareto frontier of new_pop.
    current_pop = new_pop, current_frontier = new_frontier        ▷ store current_pop
    Increase gen by 1.
end while
```

Table 1: Characterisitics of the chosen models.

Characteristics	Multi-heterotroph model (Lardon et al., 2011)	*P. gingivalis-S. gordonii* model based on (Martin et al., 2017)
Number of species	3	2
Number of nutrients	4	2
Average runtime for original model	180 minutes	50 minutes
Environment	Waste-water reactor	Human body

(restricted to 10 for time considerations). The code runs 3 (number of cores chosen) different instances of the mutated child entries concurrently. The parallelization code was based on the R package RParallel (Treadway, 2017). The average of the outputs is then taken across the repetitions and used as the fitness value. The concatenated list of the parent and children population is generated after removing repetitions. This list undergoes another non-dominated sorting and the first 20 entries are chosen to be the next parent population. All the parent populations are saved as `.csv` spreadsheets in the working directory. iDynoMiCS is run with the relevant protocol file using a call to Python from within the code.

4 Models and Results

The models used here are ABMs of bacterial biofilms. Agents are individual bacterial cells with rules for growth, division, death, extracellular polymeric substances (EPS) production, maintenance, inactivation, shoving and erosion. The growth is governed by Monod kinetics (Monod, 1949). Other state variables of the agents include location, size, density, species type and genealogy. The environment consists of a bulk compartment of well mixed solutes which diffuse into the biofilms and the surface to which the biofilms adhere. The diffusion process is ruled by partial differential equations. The computational domain is an evenly spaced rectilinear grid described by its dimensionality (2D or 3D), its size, its geometry and the behavior at its boundaries. The two selected models focus on bacterial biofilms in a waste water reactor and the human body. They were obtained from the literature (Lardon et al., 2011; Martin et al., 2017). In both papers, the parameters of the model were either extracted from experimental data or previously established literature on the specific bacteria. A reason for choosing these two particular models was ease of implementation as well as simplicity. Criteria for model selection were the ready availability of a model implementation, and limited complexity of the model, such as number of species involved, to facilitate the implementation of controls and multiple simulation runs. A summary of the two models is provided in Table 1.

For the ABMs constructed using iDynoMiCS (Lardon et al., 2011), there are some parameters that are under user control. The user can specify the initial number of agents and the region of inoculation, the initial values for each agent state variable and the size and boundaries of the computational domain. Another variable under user control is the length of the time step of the simulations. The controls we considered for both models were substrate concentrations. The outputs were the biofilm thickness (for both models) and the diversity index (for the multi-heterotroph model) or the cell count of a particular species

(for the *P. gingivalis-S. gordonii* model). For each of the models, a coarse-graining could be effected in terms of the initial cell count of the species, and the spatial dimensions or the grid/agent resolution. Multiple coarse-graining with combinations of the possibilities were carried out. The coarse-graining which contributed to the reduction of run time and a Cohen's kappa value of greater than 0.6 were chosen for the computation and comparison of the Pareto frontier. The coarse-grained models that did not match this criterion were discarded.

4.1 Multi-heterotroph model

This model represents a biofilm with three different species of bacteria (Lardon et al., 2011). The species are heterotrophs that are aerobic but can also use nitrate as an electron acceptor when oxygen is lacking. The three species are labeled according to the

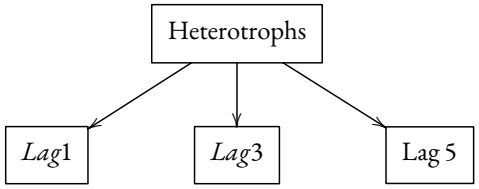

Figure 3: Species involved in multi-heterotroph model (Lardon et al., 2011).

time taken by each to activate the denitrification pathway. They are involved in the following processes:

- Growth (Aerobic and Anaerobic, using nitrate instead of oxygen)
- Maintenance (Aerobic and Anaerobic)
- Inactivation
- EPS hydrolysis

All the species have identical growth parameters. We wish to maximize the biofilm thickness and the diversity index by controlling the oxygen and nitrate concentrations. These choices are motivated by the role biofilms play in waste water treatment plants (Andersson, 2009).

For this multi-objective optimization, we first calculated the Pareto front of a two-dimensional representation after three days of growth of the biofilm in a reactor. The grid parameters are $(nI, nJ, nK) = (33, 33, 1)$, with the resolution of the computational and agent grid set at 8. In the iDynoMiCS protocol file, for a two-dimensional model, one has to set $nK = 1$. The runtime of the original model is 3 hours. The simulations were started with 10 cells of each species. Using Algorithm 1 with 50 generations and a parent size of 20, the computational time was more than 2 weeks. To reduce the run time, we changed the following model parameters: grid size, resolution, and the number of initial agents.

4.1.1 Coarse-graining of the model

To calculate the Cohen's kappa value, a list of 100 possible combinations of the oxygen and nitrate concentrations is generated and the biofilm thickness of the original model is calculated for each of these combinations. Using the same inputs, the biofilm thickness of the reduced models is also generated. These two ordinal lists are used to calculate the Cohen's kappa value.

The grid dimensions are $(nI, nJ, nK) = (33, 33, 1)$. The size and boundaries of the computational domain as well as the initial number of agents are changed. We observed the results in Table 2, which shows that coarse-graining of the model led to significant decrease in computational time and a Cohen's kappa value of more than 0.6.

The Pareto frontier of the original and coarse-grained models are depicted in Figure 4 and solute controls are shown in Figure 5.

As mentioned before, we are maximizing the biofilm thickness and the diversity index. So the Pareto front will be in the upper and right half of plot. As Figure 4 indicates, by reducing only the initial number of agents, we see that the Pareto front of the original and the reduced model (original and reduced_init in the legend) are very close to each other. On reducing both the initial count and the resolution (reduced_init_resolution1 and reduced_init_resolution2 in the legend) we see that the Pareto fronts obtained may not be close to the original but may capture points on the Pareto front that were not captured by the original model. Figure 5 shows the controls of the Pareto front, i.e., the oxygen and nitrate concentrations. We can see that the controls for the original and coarse-grained models are very similar, which is a validation of our hypothesis that the coarse-grained model works in a fashion similar to the original model.

Table 2: Run times and Cohen's kappa values of original and coarse-grained models for multi-heterotroph model. The models are original model, the model with a reduced initial seed compared to the original (Reduced_init), model with a reduction with resepect to resolution and initial seed (Reduced_init_resolution1, Reduced_init_resolution2).

	Initial Number	Resolution size: computational grid/agent grid	Run time (minutes)	Cohen's Kappa
original	10	8/8	180	
reduced_init	5	8/8	60	0.907
Reduced_init_resolution1	5	16/8	46.73	0.659
Reduced_init_resolution2	5	16/4	29.04	0.771

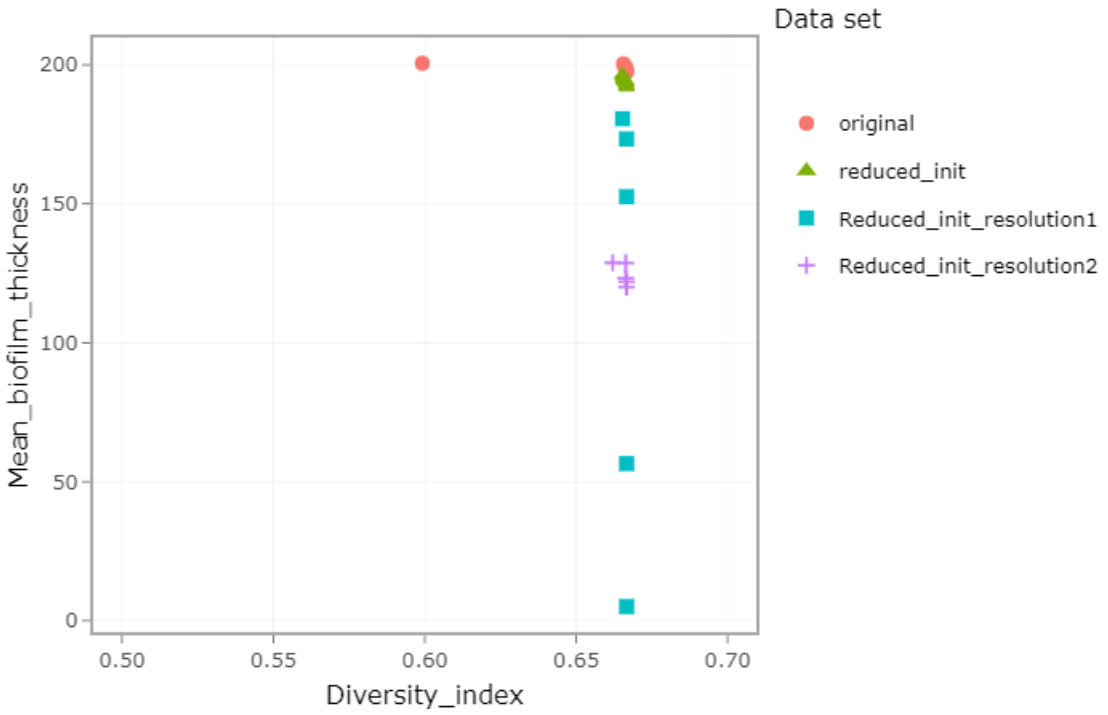

Figure 4: The Pareto front of original and coarse-grained models. In the legend, original is the original model, reduced_init is the model with a reduced initial seed compared to the original, Reduced_init_resolution1 and Reduced_init_resolution2 are the models with reduced dimensions and initial seed. See Table 2 for more details.

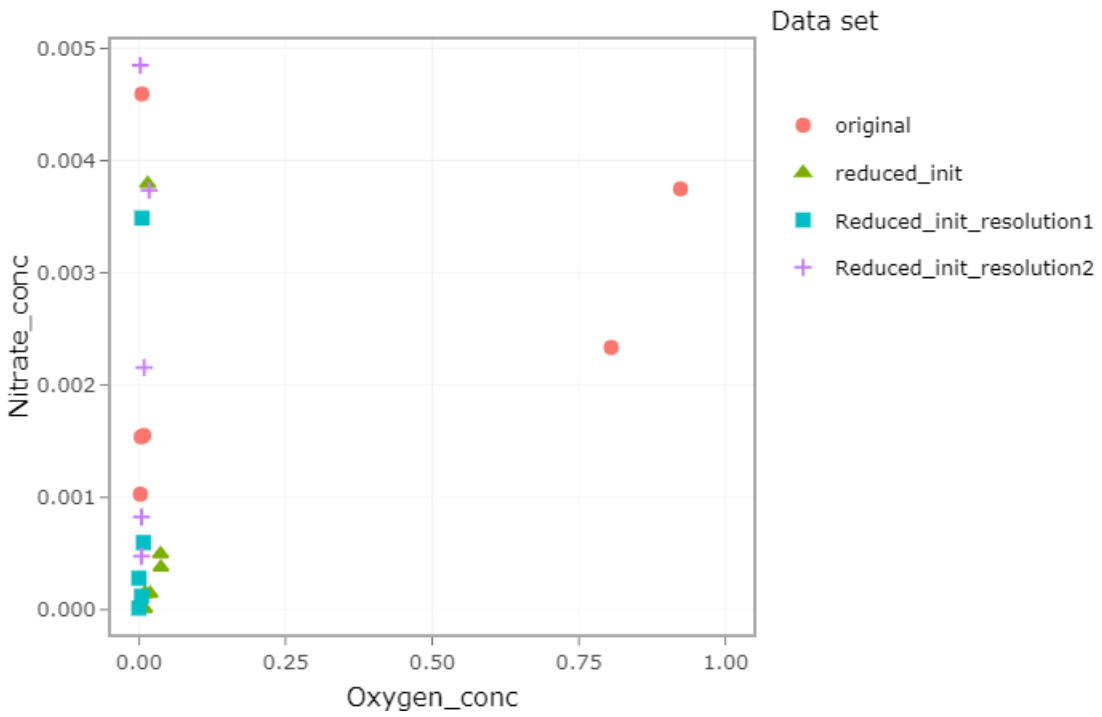

Figure 5: The controls of the Pareto front of original and coarse-grained models. The unit of concentrations in the x- and y-axes is g/L. In the legend, original is the original model, reduced_init is the model with a reduced initial seed compared to the original, Reduced_init_resolution1 and Reduced_init_resolution2 are the models with reduced dimensions and initial seed. See Table 2 for more details.

4.2 Porphyromonas gingivalis and Streptococcus gordonii model

The PDE model from (Martin et al., 2017) is a two-species model with the bacteria *Porphyromonas gingivalis* and *Streptococcus gordonii* in the biofilm. The substrate for growth of both species is protein. The model also involves the production of a substance by *S. gordonii* which is toxic to *P. gingivalis*. The model parameters available in the paper were used to create a two-dimensional ABM in iDynoMiCS which simulates a one-day growth of a biofilm. The protocol file was populated with all the relevant parameters from (Martin et al., 2017).

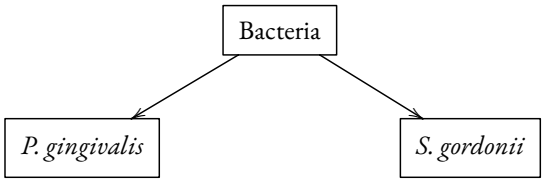

Figure 6: Species involved in *P. gingivalis* and *S. gordonii* model (Martin et al., 2017).

This model simulates a biofilm found in the body that can be pathogenic. The objective functions are to minimize biofilm thickness and the cell count of *S. gordonii* bacteria so as to reduce the toxicity of the biofilm. The controls are the protein and toxin concentrations. The grid dimensions of the original model are $(nI, nJ, nK) = (129, 129, 1)$ and the computational and agent grid resolution is 8. In the iDynoMiCS protocol file, for a two-dimensional model, one has to set $nK = 1$. The size of the grid is similar to the one employed in (Martin et al., 2017). The initial cell count is 20 and 400 for *P. gingivalis* and *S. gordonii*, respectively. Here too, to reduce computation time, as well as to maintain the same controls in the coarse-grained model, coarse-graining was carried out in terms of initial cell count, grid size or resolution.

4.2.1 Coarse-graining

To calculate the Cohen's kappa value, a list of 150 possible combinations of the protein and toxin concentrations is generated and the biofilm thickness of the original model is calculated for each of these combinations. Using the same inputs, the biofilm

Table 3: Cohen's kappa values of original and coarse-grained models for *P. gingivalis* and *S. gordonii* interactions. The models are original model, the model with a reduced initial seed compared to the original (Reduced_init), model with a reduction with resepect to resolution and initial seed (Reduced_init_resolution1).

	(nI, nJ, nK)	Resolution size: computational grid/agent grid	Initial Count	Run time (minutes)	Cohen's kappa w.r.t thickness
Original	(129, 129, 1)	8 and 8	20/400	50.7	
Reduced_init	(129, 129, 1)	8 and 8	10/200	21.99	0.912
Reduced_init_resolution1	(65, 65, 1)	8 and 8	5/100	1.59	0.807

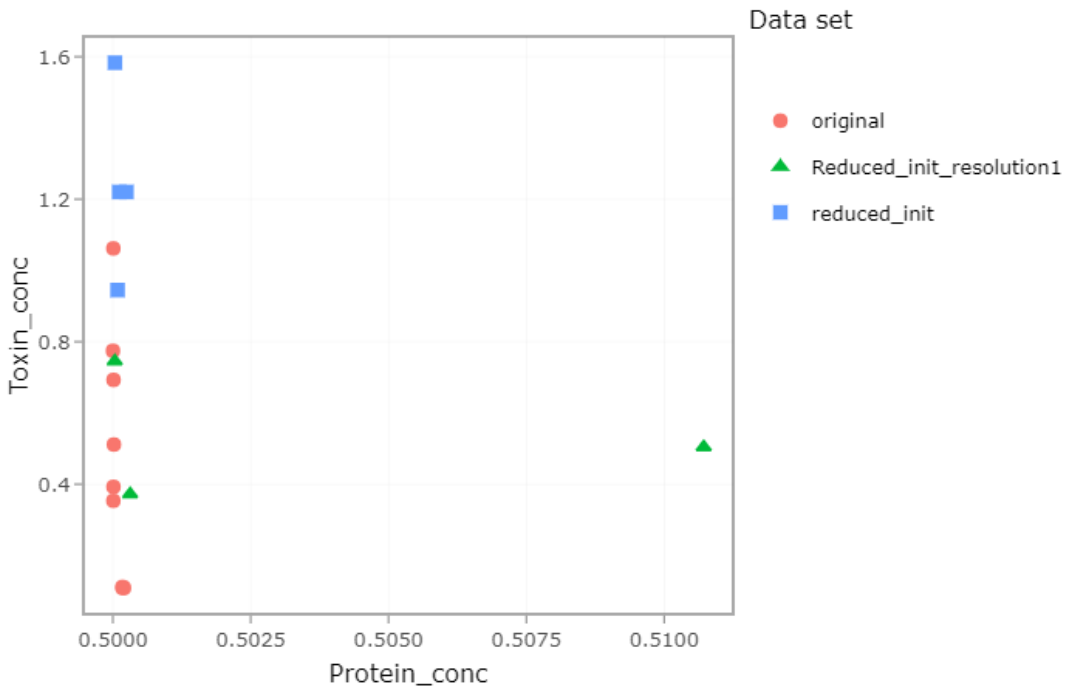

Figure 7: The controls of the Pareto front of original and coarse-grained model. The unit of concentrations in the x- and y-axes is g/L. In the legend, original is the original model, Reduced_init is the model with a reduced initial seed compared to the original, Reduced_init_resolution1 is the model with a reduction with resepect to resolution and initial seed. See Table 3 for more details.

thickness of the coarse-grained model is also generated. These two ordinal lists are used to calculate the Cohen's kappa value.

The controls for the Pareto fronts of the original and coarse-grained models are shown in Figure 7. As can be seen in Table 3, the coarse-grained models run significantly faster than the original ones and thus could be an alternative to the computationally expensive original model. Also, the controls for both the original and coarse-grained models of the Pareto front are similar (Figure 7), and this justifies the use of the coarse-grained in place of the original model.

5 Discussion

There is a clear role for mathematical modeling as the foundation for the principled development of optimal control approaches. Agent-based modeling has a long tradition as an important tool in ecology generally, and the same is true for microbial communities such as biofilms in the human body and healthcare settings. Due to limitations of ABMs discussed earlier, using them effectively for model-based control is an unsolved problem. Typically, optimal control approaches are discovered through heuristic search algorithms that require extensive model simulation. For many more complex models this is not practical. Hence model reduction of some sort becomes necessary. Coarse-graining is one such method. The novel approach proposed in (An et al., 2017) is to make the model reduction dependent on the particular control objectives to be satisfied rather than the overall similarity of the model and its reduction. This manuscript reports on an experimental study that contributes further evidence that

a combination of control-dependent coarse-graining and multi-objective control methods is an effective heuristic approach that is applicable in realistic scenarios involving multi-species bacterial biofilms.

The two models we investigate are sufficiently complex to be realistic, but are also fairly expensive to simulate, to a degree that makes heuristic control algorithms impractical to apply, except in a high-performance computing environment. Our computational experiments show that control-dependent model coarse-graining can be simulated much more quickly while retaining the effect of controls on simulation outcomes. Our study provides further evidence that this approach is promising as a practical tool for model-based control using ABMs. The novelty of this work is that the ABMs considered are more computationally and biologically complex than the models previously studied in this context.

The main limitation of this work is that it considers only two models, due to the limitated computational resources available. A more extensive study would include a broader range of models that could possibly allow the discovery of a correlation between particular model features and algorithm performance. It would also include additional control methods, mentioned in the background section, to allow benchmarking of the different methods. The results would be helpful to a practioner in deciding which particular method to use in which setting.

A further limitation is the relatively small grid size used here. To simulate real world experiments, a larger grid size as well as a larger number of agents would be needed, all of which require substantially larger computational resources than we had available. The models in this case study were two-dimensional and therefore further work will need to be done on the three-dimensional simulations of the same models. It is to be investigated whether the controls obtained in the Pareto front of the original and reduced 3D versions are comparable.

The ultimate goal of an experimental program addressing model-based control using ABMs needs to be to assess the range of models for which coarse-graining is useful and with what parameters. That is, we need to identify features of ABMs relevant for a particular optimization problem that are sensitive to coarse-graining and those that are not. This would provide at least heuristic guidelines for when to use this methodology. Since both our models are amenable to the method, we have unfortunately not learned anything useful to answer that identification problem. A much larger study would need to be carried out that would begin to allow a classification of ABMs from this point of view. In such a study, sufficiently many models related to a particular type of biological system should be chosen to make fairly broad statements. That is, it might not be easy to compare the properties of a biofilm model with those of a tumor growth model.

Acknowledgments

This work was supported by NIH Grants 1R01GM127909-01 and 3R01 GM127909-01S1. RL was also partially supported by NIH Grants 1R011AI135128-01 and 1U01EB024501-01, and NSF Grant CBET-1750183. A special thanks to Dr. Linda Archambault and Dr. Anna Dongari-Bagtzoglou for their edits.

References

An, G., B. Fitzpatrick, S. Christley, P. Federico, A. Kanarek, R. M. Neilan, M. Oremland, R. Salinas, R. Laubenbacher, and S. Lenhart (2017). Optimization and control of agent-based models in biology: a perspective. *Bulletin of Mathematical Biology 79*(1), 63–87. 167, 175

An, G., Q. Mi, J. Dutta-Moscato, and Y. Vodovotz (2009). Agent-based models in translational systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine 1*(2), 159–171. 167

Andersson, S. (2009). *Characterization of bacterial biofilms for wastewater treatment*. Ph. D. thesis, Kungliga Tekniska Högskolan. 168, 172

Bauer, E., J. Zimmermann, F. Baldini, I. Thiele, and C. Kaleta (2017). BacArena: Individual-based metabolic modeling of heterogeneous microbes in complex communities. *PLoS Computational Biology 13*(5), e1005544. 169

Bijak, J., J. Hilton, E. Silverman, and V. D. Cao (2013). Reforging the wedding ring: exploring a semi-artificial model of population for the united kingdom with gaussian process emulators. *Demographic Research 29*, 729–766. 169

Bryers, J. D. (2008). Medical biofilms. *Biotechnology and Bioengineering 100*(1), 1–18. 168

Carnell, R. (2020). *lhs: Latin Hypercube Samples*. R package version 1.0.2. 170

Chao, D. L., M. E. Halloran, V. J. Obenchain, and I. M. Longini Jr (2010). FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology 6*(1), e1000656. 169

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46. 169

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin 70*(4), 213. 169, 170

Couzin-Frankel, J. (2013). Cancer immunotherapy. *Science 342*(6165), 1432–3. 167

Deb, K., A. Pratap, S. Agarwal, and T. Meyarivan (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation 6*(2), 182–197. 170

Forhead, H. and N. Huynh (2018). Review of modelling air pollution from traffic at street-level — the state of the science. *Environmental Pollution 241*, 775–786. 167

Gamer, M., J. Lemon, I. Fellows, and P. Singh (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1. 170

Giabbanelli, P. and R. Crutzen (2017). Using agent-based models to develop public policy about food behaviours: future directions and recommendations. *Computational and Mathematical Methods in Medicine 2017*. 167

Grefenstette, J. J., S. T. Brown, R. Rosenfeld, J. DePasse, N. T. Stone, P. C. Cooley, W. D. Wheaton, A. Fyshe, D. D. Galloway, A. Sriram, et al. (2013). FRED (a framework for reconstructing epidemic dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health 13*(1), 1–14. 169

Heard, D., G. V. Bobashev, and R. J. Morris (2014). Reducing the complexity of an agent-based local heroin market model. *PloS One 9*(7), e102263. 167, 169

Heard, D., G. Dent, T. Schifeling, and D. Banks (2015). Agent-based models and microsimulation. *Annual Review of Statistics and Its Application 2*, 259–272. 169

Lardon, L. A., B. V. Merkey, S. Martins, A. Dötsch, C. Picioreanu, J.-U. Kreft, and B. F. Smets (2011). iDynoMiCS: next-generation individual-based modelling of biofilms. *Environmental Microbiology 13*(9), 2416–2434. 167, 169, 171, 172

Lazarova, V. and J. Manem (1995). Biofilm characterization and activity analysis in water and wastewater treatment. *Water Research 29*(10), 2227–2245. 168

Lee, J.-S., T. Filatova, A. Ligmann-Zielinska, B. Hassani-Mahmooei, F. Stonedahl, I. Lorscheid, A. Voinov, J. G. Polhill, Z. Sun, and D. C. Parker (2015). The complexities of agent-based modeling output analysis. *Journal of Artificial Societies and Social Simulation 18*(4), 4. 169

Li, B., D. Taniguchi, J. P. Gedara, V. Gogulancea, R. Gonzalez-Cabaleiro, J. Chen, A. S. McGough, I. D. Ofiteru, T. P. Curtis, and P. Zuliani (2019). NUFEB: A massively parallel simulator for individual-based modelling of microbial communities. *PLoS Computational Biology 15*(12), e1007125. 167, 169

Martin, B., Z. Tamanai-Shacoori, J. Bronsard, F. Ginguené, V. Meuric, F. Mahé, and M. Bonnaure-Mallet (2017). A new mathematical model of bacterial interactions in two-species oral biofilms. *PloS One 12*(3), e0173153. 171, 174

Monod, J. (1949). The growth of bacterial cultures. *Annual Review of Microbiology 3*(1), 371–394. 171

Moore, H. (1897). Cours d'économie politique. by vilfredo pareto, professeur à l'université de lausanne. vol. i. pp. 430. i896. vol. ii. pp. 426. i897. lausanne: F. rouge. *The Annals of the American Academy of Political and Social Science 9*(3), 128–131. 168

Nealon, J. and A. Moreno (2003). Agent-based applications in health care. In *Applications of software agent technology in the health care domain*, pp. 3–18. Springer. 167

Oremland, M. and R. Laubenbacher (2015). Optimal harvesting for a predator-prey agent-based model using difference equations. *Bulletin of Mathematical Biology 77*(3), 434–459. 167, 168, 170

Oyebamiji, O., D. Wilkinson, P. Jayathilake, T. Curtis, S. Rushton, B. Li, and P. Gupta (2017). Gaussian process emulation of an individual-based model simulation of microbial communities. *Journal of Computational Science 22*, 69 – 84. 169

Oyebamiji, O., D. Wilkinson, B. Li, P. Jayathilake, P. Zuliani, and T. Curtis (2019). Bayesian emulation and calibration of an individual-based model of microbial communities. *Journal of Computational Science 30*, 194 – 208. 169

Rhodes, D. M., M. Holcombe, and E. E. Qwarnstrom (2016). Reducing complexity in an agent based reaction model—benefits and limitations of simplifications in relation to run time and system level output. *Biosystems 147*, 21–27. 169

Roocks, P. (2016, dec). Computing Pareto frontiers and database preferences with the rpref package. *The R Journal 8*(2), 393–404. 170

Shreiner, A., J. Kao, and V. Young (2015). The gut microbiome in health and in disease. *Current Opinion in Gastroenterology 31*(1), 69. 167

Small, B. G., B. W. McColl, R. Allmendinger, J. Pahle, G. López-Castejón, N. J. Rothwell, J. Knowles, P. Mendes, D. Brough, and D. B. Kell (2011). Efficient discovery of anti-inflammatory small-molecule combinations using evolutionary computing. *Nature Chemical Biology 7*(12), 902. 170

Tinker, R. and U. Wilensky (2007). Netlogo climate change model. *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston*. 167

Treadway, A. (2017). Running R code in parallel. last accessed September 15, 2020. 171

Van Veldhuizen, D. A. and G. B. Lamont (1998). Evolutionary computation and convergence to a Pareto front. In *Late breaking papers at the genetic programming 1998 conference*, pp. 221–228. Citeseer. 170

Wang, Q. and T. Zhang (2010). Review of mathematical models for biofilms. *Solid State Communications 150*(21), 1009 – 1022. 168

Willem, L., S. Stijven, E. Tijskens, P. Beutels, N. Hens, and J. Broeckhove (2015). Optimizing agent-based transmission models for infectious diseases. *BMC Bioinformatics 16*. 169