PLOS Computational Biology

RESEARCH ARTICLE

Enhancing generalizability of model discovery across parameter space with multi-experiment equation learning for biological systems

Maria-Veronica Ciocanel[1☯*], **John T. Nardini**[2☯*], **Kevin B. Flores**[3*], **Erica M. Rutter**[4], **Suzanne S. Sindi**[4], **Alexandria Volkening**[5]

1 Departments of Mathematics and Biology, Duke University, Durham, North Carolin, United States of America, **2** Department of Mathematics and Statistics, The College of New Jersey, Ewing, New Jersey, United States of America, **3** Department of Mathematics, Center for Research in Scientific Computation, North Carolina State University, Raleigh, North Carolina, United States of America, **4** Department of Applied Mathematics, University of California Merced, Merced, California, United States of America, **5** Department of Mathematics, Purdue University, West Lafayette, Indiana, United States of America

☯ These authors contributed equally to this work.
* veronica.ciocanel@duke.edu (M-VC); nardinij@tcnj.edu (JTN); kbflores@ncsu.edu (KBF)

Check for updates

OPEN ACCESS

Citation: Ciocanel M-V, Nardini JT, Flores KB, Rutter EM, Sindi SS, Volkening A (2026) Enhancing generalizability of model discovery across parameter space with multi-experiment equation learning for biological systems. PLoS Comput Biol 22(4): e1014161. https://doi.org/10.1371/journal.pcbi.1014161

Editor: Alejandro F. Villaverde, Universidade de Vigo, SPAIN

Received: August 11, 2025

Accepted: March 24, 2026

Published: April 22, 2026

Copyright: © 2026 Ciocanel et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: All code and data for this study are publicly available at https://github.com/johnnardini/ME-EQL.

Funding: o All authors received travel funding for an AIM SQuaRE Program at the American Institute for Mathematics. KBF was funded in

Abstract

Agent-based modeling (ABM) is a powerful tool for understanding self-organizing biological systems, but it is computationally intensive and often not analytically tractable. Equation learning (EQL) methods can derive continuum models from ABM data, but they typically require extensive simulations for each parameter set, raising concerns about generalizability. In this work, we extend EQL to Multi-experiment equation learning (ME-EQL) by introducing two methods: (i) one-at-a-time ME-EQL (OAT ME-EQL), which learns individual models for each parameter set and connects them via interpolation, and (ii) embedded structure ME-EQL (ES ME-EQL), which builds a unified model library across parameters. We demonstrate these methods by learning continuum models from a noisy birth–death mean-field model and from an on-lattice agent-based model of birth, death, and migration with spatial structure, often used to investigate cell biology experiments. We show that both methods significantly reduce the relative error in recovering parameters from agent-based simulations, with OAT ME-EQL offering better generalizability across parameter space. Our findings highlight the potential of equation learning from multiple experiments to enhance the generalizability and interpretability of learned models for complex biological systems.

Author summary

Biological systems often display complex patterns and dynamics across space and time in response to interactions between individual units, such as cells, molecules, or animals. Mathematical modeling is an essential tool to understand how biological interactions scale into emergent behaviors. Agent-based models

UNCORRECTED PROOF

part by the National Science Foundation under grant numbers DMS-2327836, DMS-2342344, and DMS-2424748. KBF and MVC were funded in part by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health under grant number 1U54AI191253-01. The funders did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

are an especially powerful framework for investigating relevant biological mechanisms by simulating interactions between agents. A limitation of these models, however, is their computationally intensive nature and the large number of input parameters. These challenges hinder modelers' ability to connect agent-based models to biological data for data-driven tasks such as parameter inference, which requires numerous model simulations. Here, we propose novel extensions of sparse regression techniques for equation learning, aimed at learning parameterized differential equations from multiple agent-based model experiments. Our methods derive differential equations that describe the population dynamics across parameter regimes, which we demonstrate using a spatial birth–death–migration model commonly applied to cell biology experiments. We find that these data-driven methods learn equations that generalize across parameter space and also significantly improve the recovery of parameters from agent-based model datasets.

Introduction

Biological systems can exhibit rich spatiotemporal patterns and dynamics in which population-level behavior emerges from interactions at the level of individual entities, i.e., cells, organisms, molecules and animals [1–4]. Mathematical models can be used as quantitative tools for bridging our understanding of how interactions across multiple scales lead to emergent behaviors. Mechanistic mathematical models provide a powerful framework for investigating these biological systems, with model complexity and abstraction tailored to the biological question and computational constraints [5–7]. Among mechanistic approaches, agent-based models (ABMs) are characterized by their high level of detail, i.e., with the capability to investigate biologically-relevant mechanisms and capture spatial effects by explicitly simulating individual agents and their interactions. However, the computational demands of ABMs often limit their utility in large-scale inference tasks such as parameter estimation, uncertainty quantification, sensitivity analysis, and optimal design. To overcome these limitations, surrogate models consider aggregate or coarse-grained agent dynamics, enabling efficient computational exploration of parameter spaces while preserving key features that are causal to the underlying biological dynamics [8–12].

The task of finding accurate surrogate models is challenging. Parameterized differential equation models are often preferable to "black-box" models, such as neural networks or Gaussian processes, because they offer interpretability, enable symbolic analytical techniques, and are compatible with established methods for parameter estimation and uncertainty quantification. Analytical methods exist to formally derive mean-field models from ABM simulations, such as moment closure methods [13–15] and the Fokker-Planck equation [16]. However, while mean-field approaches offer analytical and computational tractability, they generally rely on strong formal assumptions, such as the diminishing effect of higher-order correlations or homogeneous mixing. These assumptions may not hold for many ABMs of biological systems, which

UNCORRECTED PROOF

exhibit spatial structure, stochasticity, and heterogeneous interactions, making analytical derivations difficult to apply and validate [17,18]. Moreover, the behavior of ABMs can vary significantly across parameter regimes, with spatial correlation effects emerging or becoming negligible depending on the specific parameter values. This variability motivates the need for frameworks to derive differential equation surrogate models that can robustly capture essential dynamics across a broad range of ABM parameter regimes [10,19].

Equation learning (EQL) has emerged as a powerful tool for using time-series data to discover governing differential equations by identifying functions that accurately describe the rate of change of biological processes driving the system dynamics. EQL methods aim to solve the problem of symbolic regression, in which regression methods are used to find the mathematical expressions that best describe a model for fitting data. Symbolic regression encompasses genetic programming methods [20,21] and neural network or statistical learning approaches [22,23], among others. For a recent review on the topic, see [24]. One prominent approach in this field is the Sparse Identification of Nonlinear Dynamical Systems (SINDy) method, which relies on sparse regression techniques to identify the underlying equations governing complex systems [22]. While approaches like SINDy have predominantly been applied to uncovering governing equations of dynamical systems, their potential is expanding into other fields, including biology and ecology, where similar challenges in system identification exist [25,26]. Recently, Nardini et al. proposed using EQL to automatically derive ODEs as surrogate models for ABMs of biological phenomena [18]. While this approach successfully demonstrates the feasibility of learning interpretable mean-field models from ABM simulations, it was applied in a restricted setting in which surrogate models were only derived for fixed values of ABM parameters. In particular, each new parameter set requires retraining the surrogate model. This limitation restricts the ability of EQL to generalize across parameter spaces where emergent behaviors may change qualitatively. Recent advances in conditional equation learning and operator learning aim to address this by capturing parameter-to-dynamics mappings, but these methods often sacrifice interpretability or require latent representations that are difficult to interpret biologically [27]. These gaps highlight the need for new methods that can derive differential equations from ABMs in a way that generalizes across parameter regimes while preserving mechanistic insight.

Here we develop two methods for learning generalizable differential equations from multiple ABM simulations conducted across different parameter values (i.e., a set of computational experiments). We refer to our methods collectively as multi-experiment equation learning or "ME-EQL" (Fig 1). We consider ME-EQL approaches, including the two proposed here, as equation learning methods that also serve as a form of "model discovery," because they recover explicit governing equations representing a unified mechanistic model shared across parameter regimes. Our first approach, which we refer to as embedded structure multi-experiment EQL (ES ME-EQL), includes the biological parameters being varied explicitly as coefficients in the function library. Sparse regression is then applied to experiments from all observed parameters at the same time. In our second method, which we refer to as one-at-a-time multi-experiment EQL (OAT ME-EQL), we perform equation learning for each parameter choice separately and then interpolate coefficients over experiments to produce a model that generalizes across parameter space.

Overall, we find that both ME-EQL methods exhibit significant promise for learning from a birth–death–migration model on a spatial lattice. This simple ABM is a canonical biological model, where the agents can represent cells in wound healing [28] or animal movement in ecology [29]. From this ABM, one can derive equations that have been broadly used across biological systems, e.g., the Fisher-KPP equation [30–32]. Moreover, this ABM and its extensions have previously been used to assess EQL performance [10,18]. The success of these methods demonstrates the ability of learned differential equations to capture unmodeled effects such as spatial correlations and interactions between neighboring biological agents not explicitly included in the EQL library construction. We are not the first to consider the problem of learning across parameter space, as other studies have considered embedding the model parameters in a similar way [33]. However, to our knowledge, the work presented here is the first to compare different ME-EQL approaches and consider agent-based biological dynamics. This development of ME-EQL methodology is a prerequisite to applying EQL to more realistic biological scenarios in which no single differential equation form captures the full diversity of experimental conditions [34].

UNCORRECTED PROOF

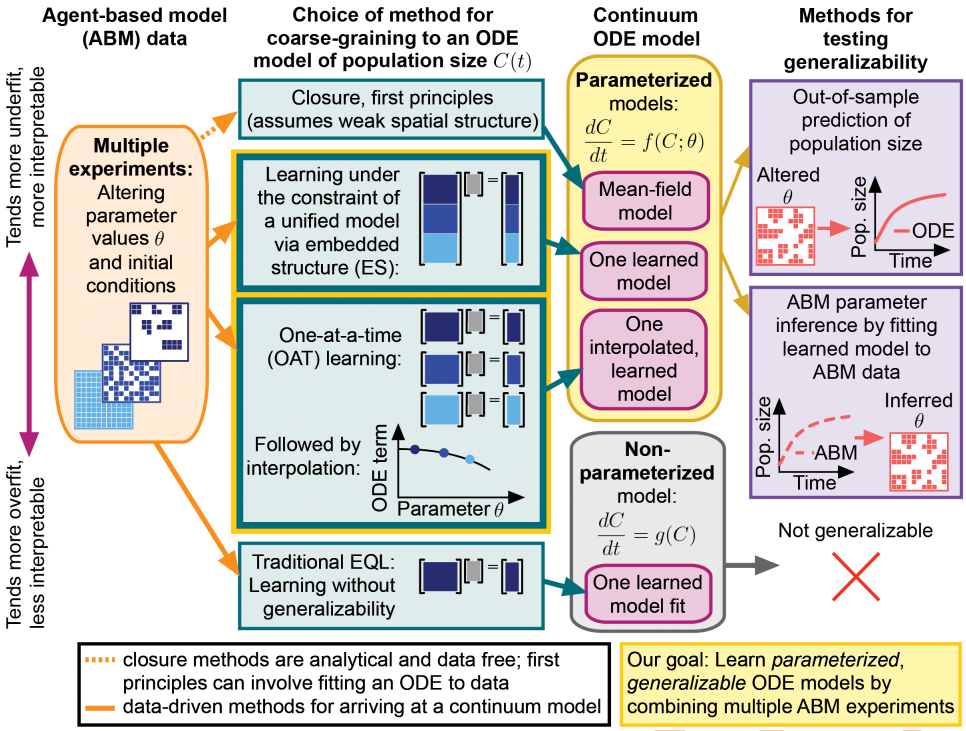

Fig 1. Overview of our motivation and approach. Agent-based models (ABMs) are a natural means of describing many biological systems, but these stochastic models often encounter challenges when researchers attempt analysis or parameter inference. Here, we describe new methods to utilize information from multiple experiments arising from different ABM parameter regimes (Orange). Traditional methods to develop coarse-grained models rely on closure assumptions that may lead to inaccurate representations of ABM spatial structure (Light Green, Top). Alternatively, traditional equation learning (EQL) methods involve discovering models from data, leading to excellent fits on training data but no means of generalizing to out-of-sample prediction (Light Green, Bottom). We propose two methods (Yellow) for addressing these challenges by performing EQL from multiple ABM experiments under different parameter values: ME-EQL. Our first method, ES ME-EQL, relies on learning ODEs from a library with embedded structure (ES) in the form of data and parameters from multiple ABM simulations; our second method, OAT ME-EQL, consists of repeating traditional EQL with different ABM parameters followed by interpolation to map these models to unobserved parameter values. Our approaches lead to parameterized ODEs (Pink), and we test their generalizability and interpretability by predicting ABM population size and inferring ABM parameter values (Purple).

https://doi.org/10.1371/journal.pcbi.1014161.g001

Methods and experiments

Generating data

Mean-field model. Following [18], we used a classical birth–death (BD) model, a fundamental example in mathematical biology that describes a well-mixed population for an ABM on a lattice where individuals undergo birth, death and migration events. Agents proliferate (giving rise to a new agent) with rate R_p and die (and are removed from the lattice) with rate R_d. Invoking the mean-field assumption in this ABM produces a mean-field differential equation model for $C_{MFM}(t)$, which describes the evolution of the population density over time:

$$\frac{d}{dt}C_{MFM}(t) = R_p C_{MFM}(t)\left(1 - C_{MFM}(t)\right) - R_d C_{MFM}(t) \tag{1}$$

$$= (R_p/2)C_{MFM}(t) - R_p C^2_{MFM}(t). \tag{2}$$

UNCORRECTED PROOF

To simplify our analysis, we set $R_d = R_p/2$ and vary $R_p = \{0.01, 0.02, \ldots, 5\}$. We consider two initial conditions: $C_{MFM}(0) = 0.05$ and $C_{MFM}(0) = 0.25$, which we denote IC = 0.05 and IC = 0.25, respectively.

We simulate data $C_d(t)$ under differing amounts of noise:

$$C_d(t) = C_{MFM}(t) + \overline{C_{MFM}(t)}\mathcal{E} \tag{3}$$

where $\overline{C_{MFM}(t)}$ represents the mean value of $C_{MFM}(t)$ and $\mathcal{E} \sim N(0, \sigma)$ represents i.i.d. Gaussian noise with standard deviation σ. We examine two cases of the MFM under differing noise levels: (1) No noise ($\sigma = 0$) and (2) low noise ($\sigma = 0.0025$).

Agent-based model We generated synthetic data from an ABM for the scenario where no differential equation model is known to accurately approximate the ABM for all parameters. In the ABM, each agent exists on a two-dimensional lattice with reflecting boundary conditions. We assume that each agent can proliferate with rate R_p, dies with rate $R_d = \frac{R_p}{2}$, and migrates to an adjacent lattice site at the rate $R_m = 1$. Unlike the MFM model, the ABM captures spatially-heterogeneous interactions, but this spatial structure is *hidden* in our observations because we only track the population density over time. Moreover, this ABM has been previously shown to be approximated using the same mean-field model as Eq (2) in some parameter regimes [17,18]. This allows us to compare the learned equations in different parameter regimes, where the mean-field assumption may or may not be accurate. For further details about simulating the ABM, see [18].

Each simulation is independently initialized by selecting 5% or 25% of the lattice sites uniformly at random for occupation. We denote these initial conditions as IC = 0.05 and IC = 0.25, respectively. For each choice of proliferation rate R_p, we generate 25 independent simulations, and for each simulation we track the total density of occupied sites over time:

$$C^{(i)}_{ABM}(t) = \frac{T^{(i)}(t)}{X^2},$$

where $X^2 = 120^2$ is the size of the lattice and $T^{(i)}(t)$ is the total number of occupied sites in simulation i and at time t. The data we consider for equation learning, $C_d(t)$, is the average of all ABM simulations:

$$C_d(t) = \langle C_{ABM}(t) \rangle = \frac{1}{N}\sum_{i=1}^{N} C^{(i)}_{ABM}(t), \tag{4}$$

where $N = 25$ is the number of simulations averaged for each R_p value.

Equation learning

The goal of an equation learning (EQL) framework is to learn the dynamical systems model given by

$$\frac{dC(t)}{dt} = \mathcal{F} \tag{5}$$

that best describes observations of the dynamics, $C_d(t)$. Note that for simplicity of notation, we do not include subscripts to denote the time points at which data are observed, i.e., $C_d(t)$ is short-hand notation for data collected at times $\{t_i\}_{i=1}^n$ corresponding to $\{C_d(t_i)\}_{i=1}^n$, which might contain observation or process noise. Our EQL method builds on the SINDy (Sparse Identification of Nonlinear Dynamics) methodology [22]. The general approach is that a library of potential terms is created for $\mathcal{F}$, and sparse regression is used to select the most parsimonious model that describes the data. In the following, we discuss the steps involved in EQL: Step (1) approximating the time derivative of the data, Step (2) constructing the library, and Step (3) sparse regression for model selection.

UNCORRECTED PROOF

Step 1: Derivative approximation from data. To find the appropriate right-hand side of Eq (5), we must calculate $\frac{dC_d(t)}{dt}$ using data $C_d(t)$. Previous studies have shown that the presence of noise in the observed data can be amplified when using finite-differencing to calculate derivatives [35]. To account for this, we use `smoothdata` in Matlab to smooth the derivatives obtained using forward finite differences for our ABM data. In the case of the MFM data, we control the (small) amount of noise added to the data and numerically approximate the derivatives using the `numpy.gradient` function for central finite differences in Python.

Step 2: Library construction. The library of potential right-hand side terms is constructed by forming a matrix Θ in which the rows correspond to time points and columns correspond to library terms evaluated at those time points. In this manuscript, we use polynomial terms. However, any functional forms that the user postulates are important to explain the underlying dynamical system that generated the data could be included (e.g., trigonometric or exponential functions).

Step 3: Sparse regression. For a library of model terms Θ, and time derivatives of the data $\frac{dC_d(t)}{dt}$, sparse regression is applied to the linear equation defined by

$$\frac{dC(t)}{dt} = \Theta\xi \tag{6}$$

to estimate a sparse vector of parameters, ξ, found by solving the optimization problem

$$\hat{\xi} = \arg\min_{\xi} \left|\left| \frac{dC_d}{dt} - \Theta\xi \right|\right|_2^2 + \lambda \left|\xi\right|_1. \tag{7}$$

The ℓ_1 penalty added to the objective function promotes sparsity and the hyperparameter λ is used to tune overfitting [36]. We use cross-validation and Akaike Information Criteria (AIC) scores to select the optimal λ, as is used in existing literature [37]. While the details can be found in the hyperparameter selection section in S1 Text, we briefly describe the process below:

- Randomly split the data into 10 training and testing sets, each with 80% training and 20% validation, denoted $C_{d,k}$, where $k = 1,..,10$ denotes the train-test splits.
- For each train-test split, we solve Equation (7) using the pysindy package in python [38,39] over a grid of 100 equi-log-spaced values $\lambda_j = 10^{-1}, ..., 10^{-9}$ to obtain the optimal coefficients $\xi_{k,j}$.
- For each $\xi_{k,j}$, we forward solve Equation (6) and calculate the AIC score. At each λ_j value, we average the 10 AIC scores from the test-train splits to obtain $\bar{\lambda}_j$.
- The final λ value is selected by being the smallest $\bar{\lambda}_j$ while also satisfying that $|\xi_{j,k}|$ is below a pre-defined threshold for all 10 test-train splits. The first criteria ensures a parsimonious fit while the second criteria leverages domain knowledge that coefficients should not be overly large (see S1 Fig to visualize AIC score versus $\bar{\lambda}$).

Multi-experiment equation learning (ME-EQL). We consider two learning approaches for enhancing generalizability to parameter sets not used in the training data for equation learning. We perform the optimization problem defined by Eq (7) over data arising from multiple experiments. We refer to the first approach as "one-at-a-time" (OAT ME-EQL) (parameter-specific/individual experiment) and to the second approach as "embedded-structure" (ES ME-EQL). Step 1 (approximating the time derivative from data) is the same for each method, although the approaches diverge in Steps 2 and 3.

OAT ME-EQL. In this approach, we learn the underlying dynamics independently for each dataset generated using parameter R_p. In Step 2, the library of potential terms is $\Theta = [C, C^2, \ldots, C^{10}]$. In Step 3, the hyperparameter λ is selected for each R_p dataset independently (see hyperparameter selection section in the S1 Text for more details). In step 4, the

UNCORRECTED PROOF

threshold for $|\xi_{j,k}|$ is set to 100. This results in separate $\hat{\xi}$ learned for each R_p in Eq (7). This single-experiment learning (from each individual R_p) uses SINDy and hyperparameter selection, and we refer to it as **OAT EQL** (one-at-a-time EQL). We thus learn potentially different model structures and coefficients for the datasets corresponding to each parameter value. We note that our approach of combining information from multiple simulations shares similarities with previous ensemble-style equation learning methods, i.e., Ensemble-SINDy [40]. However, our OAT-ME-EQL method differs because it utilizes data generated from distinct parameter values, whereas Ensemble-SINDy uses simulations generated by the same fixed parameter values but with varied initial conditions. Interpolation across coefficients of the most commonly-learned right-hand-side terms is performed, yielding a single equation that generalizes across all parameters R_p. See Algorithm 1 in S2 Text and Generalizability over parameter space for more details.

ES ME-EQL. In this approach, a single model is learned for all parameter values R_p and over all datasets jointly. To accomplish this, the library of potential terms in Step 2 is assumed to be $\Theta = [R_pC, R_pC^2, \ldots, R_pC^{10}]$. In other words, we assume that the coefficients in the learned ODE are linear in R_p; this simple parameter dependence is inspired by the mean-field model (2). Similarly, the hyperparameter λ is selected jointly over all datasets to ensure that only one model is learned. In step 4, the threshold for $|\xi_{j,k}|$ is set to 20. Thus, in this case, a single $\hat{\xi}$ is learned corresponding to only one model structure that is learned for all R_p values. See Algorithm 2 in S2 Text and Generalizability over parameter space for more details.

Generalizability over parameter space

A key aim of our framework is to enable generalization to parts of parameter space that were not used to learn the differential equation (DE) models. In each of the equation learning methods we consider, hyperparameter tuning and model selection occurs using model simulations from a select number of parameter sets (e.g., 5 or 10 separate R_p values, which we refer to as experiments). Below, we describe our methodology for using learned DE models to estimate parameters from model simulations that were not included in the training data set.

OAT ME-EQL. In this framework, it is possible that different models structures are learned for each dataset corresponding to a different value of the parameter R_p. Model structure is defined as the collection of library terms with non-zero coefficients. To enable generalization to unseen parameters, we select the most common model structure given by:

$$\frac{dC}{dt} = \xi_1 C + \xi_2 C^2 + \cdots + \xi_{10} C^{10}, \tag{8}$$

which holds for the majority of R_p values that resulted in the same set of non-zero coefficients ξ_i. The coefficient vectors $\boldsymbol{\xi_i}, i = 1, \ldots, 10$, contain the parameters ξ_i learned for all R_p values. We then interpolate coefficients across parameter space using cubic splines or lines to obtain a function $\xi_i(R_p)$. We only use the most common model for interpolation. For example, if we select 10 R_p values, but only 8 have the same model structure, we would only use those 8 coefficient sets for interpolation. The generalized model is then given by:

$$\frac{dC}{dt} = \xi_1(R_p) C + \xi_2(R_p) C^2 + \cdots + \xi_{10}(R_p) C^{10}. \tag{9}$$

Note that if there is no dominant common model structure, this provides insight into whether it is possible for a single differential equation model from our library to generalize across parameter space.

ES ME-EQL. Since one unified model is learned over all parameter sets and R_p is incorporated into the library, generalization to outside parameter sets is trivial. Our learned model has the structure:

$$\frac{dC}{dt} = \xi_1 R_p C + \xi_2 R_p C^2 + \cdots + \xi_{10} R_p C^{10}. \tag{10}$$

UNCORRECTED PROOF

Thus, to predict $C(t)$, Eq (10) is simulated at a given out-of-sample parameter value of R_p.

Results

We test both ME-EQL frameworks (OAT ME-EQL and ES ME-EQL) with increasingly complex data. First, we examine the ability of the proposed algorithms to generalize across parameters when the data is generated using a known underlying model in Learning for noisy mean-field model data. We examine how the methods perform when increasing noise or decreasing information content, which we control by altering the initial condition to reduce the dynamic range of population densities observed (see S2 Fig). Secondly, we investigate algorithm performance in a situation where there may not be a known underlying model by using ABM data output in Learning for agent-based model data. Lastly, in Can we infer R_p from ABM data?, we examine the ability to recover ABM parameters from a single ABM simulation using our learned equations.

We also compare the results from both ME-EQL frameworks for these datasets with two established methods: (1) mean-field model approximations and (2) EQL for one parameter at a time, with no interpolation. Recall that we refer to this second method as OAT EQL (see Equation learning), since it finds an equation for each R_p value but is not intended to generalize its results between R_p values.

Learning for noisy mean-field model data

We first consider data generated using the mean-field model Eq (2) under no noise or low noise (σ = 0.25%). We denote this by MFM data. To assess the impact of information content on the learned model and generalizations, we also examine two initial conditions: IC = 0.05 and IC = 0.25. As we show in S2 Fig, shifting from IC = 0.05 to IC = 0.25 effectively means that we observe a much narrower range of the population dynamics, leading to less information content. (For example, notice that if we were to consider the extreme case of IC = 0.5, the mean-field model would be at equilibrium and there would be no information—other than the equilibrium value—available in the data for equation learning.) Fig 2 displays the comparison between a single MFM dataset with 0.25% noise (black stars) and the resulting OAT EQL fit (blue line) and ES ME-EQL fit (green line) for proliferation rate R_p = 0.1. These results indicate that the learned models accurately fit the simulated data.

First, we examine the models learned using our ME-EQL frameworks using data simulated at proliferation rate values $R_p = \{0.01, 0.02, \ldots, 5\}$. Fig 3a shows that there is clear agreement between the true coefficients (black lines), the learned

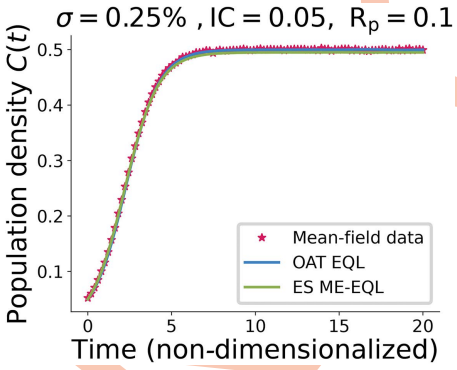

Fig 2. Sample dataset generated using the Mean-Field Model (Eq (2)) with added 0.25% proportional noise (black stars) and fits with the OAT EQL approach (blue line) and the ES ME-EQL approach (green line). The sample MFM dataset shown here is generated using proliferation rate R_p = 0.1 and initial condition IC = 0.05.

https://doi.org/10.1371/journal.pcbi.1014161.g002

UNCORRECTED PROOF

OAT EQL coefficients (circles), and the learned ES ME-EQL coefficients (squares and triangles). Over the 500 R_p values used to generate MFM data, the OAT EQL method learns the correct underlying model in 498 cases, while the other two cases learn a structure that incorporates a C^3 term (Fig 3b). The recovered underlying model for the ES ME-EQL and OAT ME-EQL approaches is the correct model (Fig 3c).

When adding small levels of noise (σ = 0.025%) to the generated data, there are clear impacts for the recovered models for both the OAT EQL and ES ME-EQL approaches. Fig 3d indicates that the learned coefficients do not always reflect the correct underlying model, as in the noise-free case. For example, in the OAT EQL recovered coefficients, there are not only C^1 (blue circles) and C^2 (orange circles) coefficients, but sometimes coefficients for $C^3, \ldots, C^6$ terms. While many of the recovered coefficients agree with the true coefficients (black line), there are some deviations. For the ES ME-EQL recovered coefficients, we learn coefficients for C^1 (squares), C^2 (triangles), and C^3 (diamonds). The C^1 coefficients match well with the underlying true coefficient values, however there are slight deviations in the C^2 coefficients, especially as R_p increases. The learned C^3 coefficients are small, but no such term exists in the underlying model. OAT EQL learns a larger variety of model structures (Fig 3e) with this small amount of added noise, but identifies the correct model for over 60% of the R_p values. ES ME-EQL thus does not learn the correct model (given the small C^3 term), while the OAT ME-EQL method most commonly learns the correct model (Fig 3f).

To assess the generalizability of the methods, we apply the learning frameworks to a smaller set of the data, generated using only 10 or 5 R_p values. When learning from 10 experiments, we select $R_p = [0.01, 0.51, 1.01, \ldots, 4.51]$ and when learning from 5 experiments, we select $R_p = [0.01, 1.01, \ldots, 4.01]$. As described in Generalizability over parameter space, we generalize the learned equations to unseen R_p values. The learned models can be found in S1 Table. We then compare the mean squared error (MSE) between the generalized recovered model and the noisy ABM data (Eq (4)) corresponding to each R_p parameter.

In the case of noise-free data, it is clear that all methods perform similarly in terms of MSE, except when R_p values are small (Fig 4a,b). Surprisingly, even with as few as 5 experiments (Fig 4a), the correct underlying model is learned for both the OAT ME-EQL and ES ME-EQL methods. The interpolation introduces little error in the OAT ME-EQL method, except when proliferation rates are small. As we increase from 5 to 10 experiments for interpolation, the region of R_p space with higher MSEs slightly decreases (Fig 4b). In the case of noisy data, both OAT ME-EQL and ES ME-EQL appear to have MSE values that are on par with the known underlying noise level (Fig 4c,d). However, the ES ME-EQL method learns an additional C^3 term. Despite this, the recovered MSE values are small. We also find that OAT ME-EQL learns the correct model even with the introduction of noise. In general, all methods recover MSEs on the same order of magnitude, resulting in similar forward solutions (see Fig 2).

We also examined the learning outcomes when using the larger initial condition of IC = 0.25 (Fig 5). The ES ME-EQL learned model contains an extra C^3 term, even when the data is noise-free (Fig 5c). However, the learned coefficients in the ES ME-EQL model are similar to the known underlying coefficients and the coefficient in front of the C^3 term is small. The learned models using OAT EQL are mostly the correct model form (>80%, Fig 5b), although higher order terms are learned for a small proportion of R_p values.

In terms of generalizability, the OAT ME-EQL approach outperforms ES ME-EQL. When using 5 experiments for interpolation (Fig 5d), the MSEs are several orders of magnitude smaller than those from ES ME-EQL. This is likely due to ES ME-EQL learning the incorrect model with slightly different coefficients (see S1 Table). We observe similar results when learning from 10 experiments (Fig 5e). Compared to the IC = 0.05 case, we find that (1) equation learning more frequently fails to capture the underlying model in the OAT EQL case, and (2) ES ME-EQL does not always recover the true model even when the data is noise-free. We suspect that this is due to the smaller information content in the data corresponding to the IC = 0.25 initial condition.

Overall, when the underlying model is known, we find that both methods perform similarly well in recovering models that match the data with small residuals. The ES ME-EQL approach has the advantage that it is required to learn a single equation,

UNCORRECTED PROOF

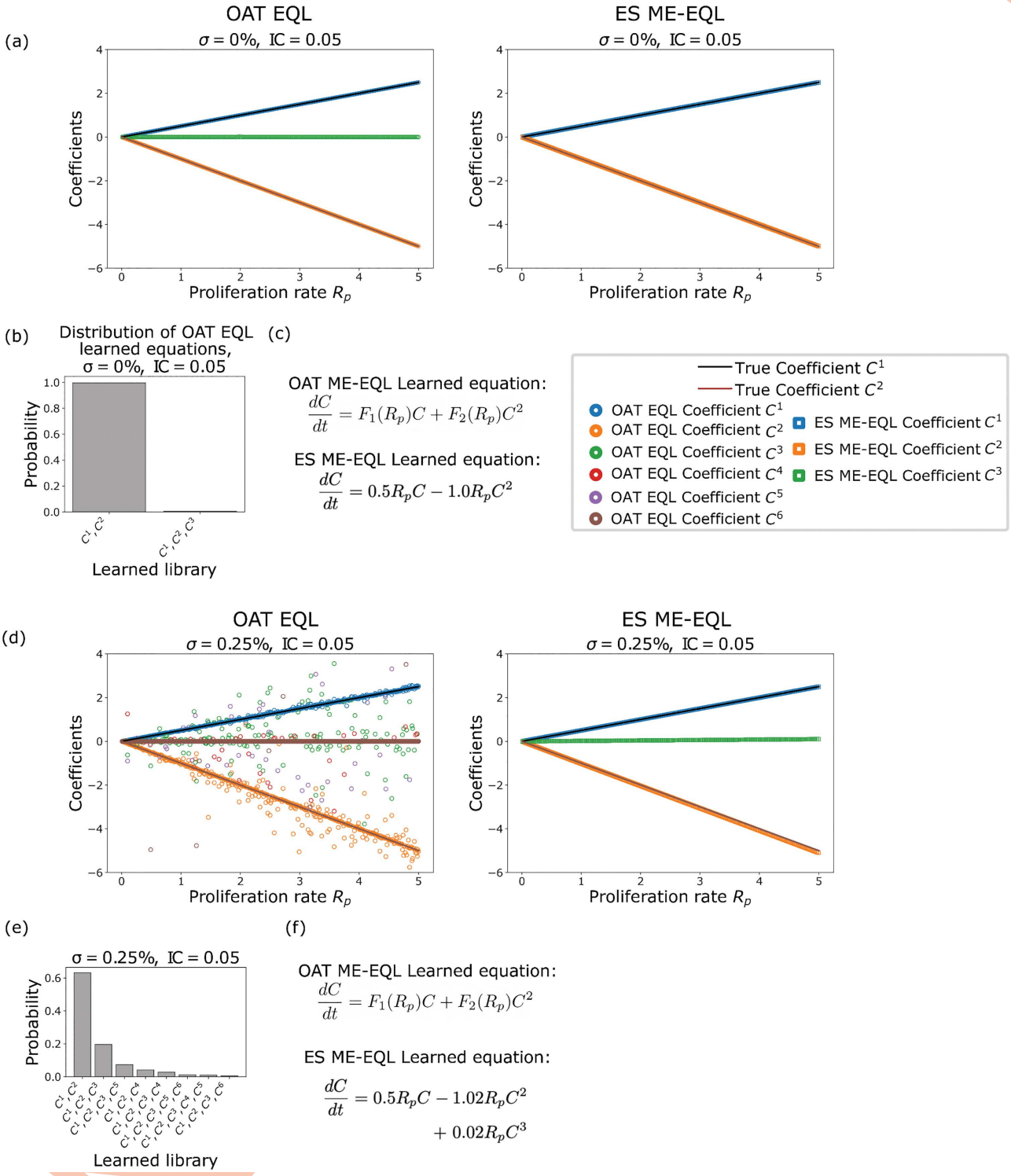

Fig 3. Learned coefficients and models for the mean-field model Eq (2) for IC = 0.05 with no noise (top) and 0.25% noise (bottom). Panels **(a)**, (d) display the true model coefficients (black lines), learned model coefficients using OAT EQL (colored circles), and learned model coefficients using ES ME-EQL (hollow shapes). Panels **(b)**, (e) depict histograms of the frequencies of the learned models for OAT EQL. Panels **(c)**, (f) list the learned OAT ME-EQL and ES ME-EQL models. In the noise-free case **(a-c)**, both OAT ME-EQL and ES ME-EQL learn the model coefficients accurately. The recovered model for ES ME-EQL and the most commonly recovered model for OAT EQL (99.6%) is the true underlying model in Eq (2). For the case

with noise **(d-f)**, the recovered coefficients do not always match the known model coefficients. However, 63% of the learned OAT EQL models recover the true underlying model Eq (2). ES ME-EQL recovers a small extra cubic term.

https://doi.org/10.1371/journal.pcbi.1014161.g003

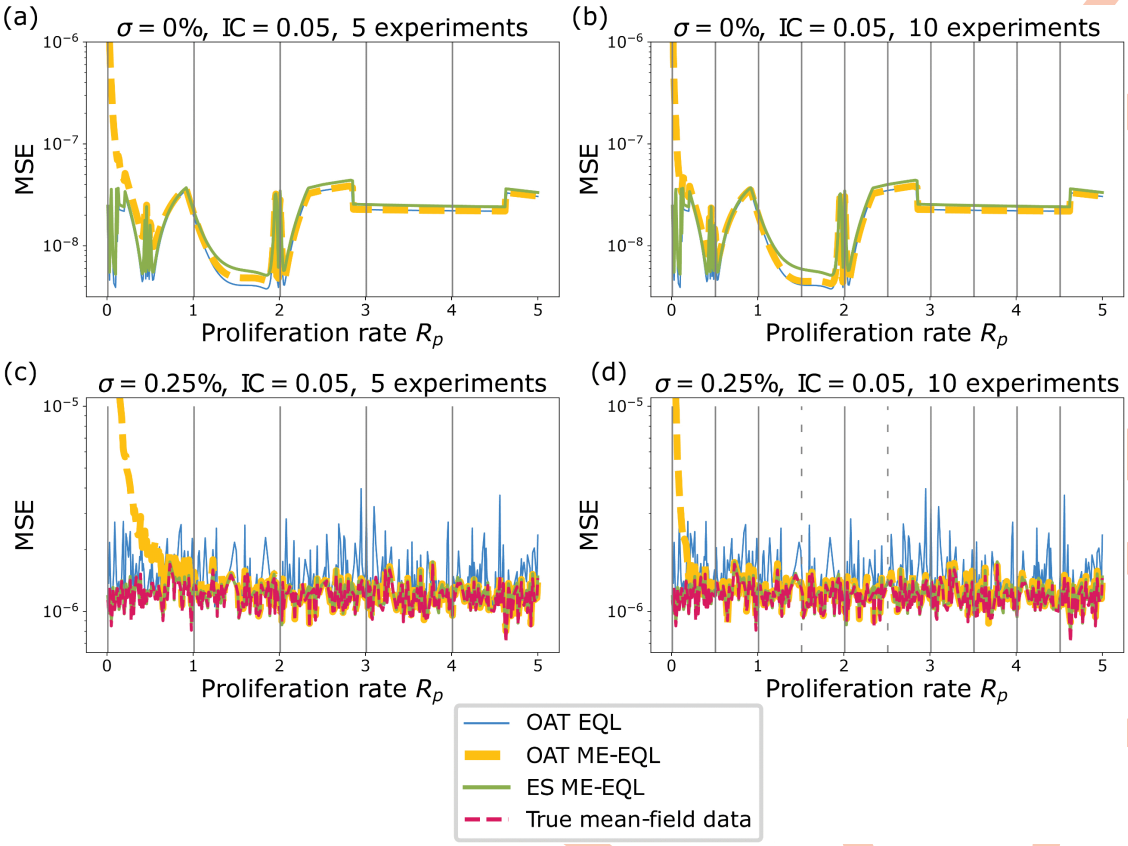

Fig 4. MSE between data and recovered models for 0% noise (a-b) and 0.25% noise (c-d). The results from OAT EQL from each separate R_p value are shown in blue for comparison purposes, the OAT ME-EQL learning is shown in yellow dashes, and the ES ME-EQL learning is shown in green solid lines. The gray vertical bars indicate the small set of R_p values from which coefficients were learned from. Dashes indicate that OAT ME-EQL did not include the dataset corresponding to that R_p value, since this framework did not learn the most popular model at that parameter value. In panels **(a)**, **(c)**, OAT ME-EQL and ES ME-EQL learn from maximum 5 R_p values, and in panels **(b)**, **(d)**, OAT ME-EQL and ES ME-EQL learn from maximum 10 R_p values. The red dashed lines represent the error added to the MFM model, which is only shown in the noisy case.

https://doi.org/10.1371/journal.pcbi.1014161.g004

which is often the true model, and performs well on out-of-sample experiments. There are, however, settings where it learns an additional term of the ODE model, especially when confronted with more noise and less information content. In contrast, the OAT ME-EQL approach uses the most popular learned equation and interpolates the coefficients across parameter values. In general, the most popular recovered model matches the known underlying dynamical system, however the interpolated coefficients do not extrapolate well to small R_p values. Both methods perform well when learning from as few as 5 experiments.

Learning for agent-based model data

To assess generalizability in cases where there is no known underlying dynamical system, we apply OAT ME-EQL and ES ME-EQL to the ABM data generated as described in Generating data. Fig 6 displays sample ABM data (black stars)

UNCORRECTED PROOF

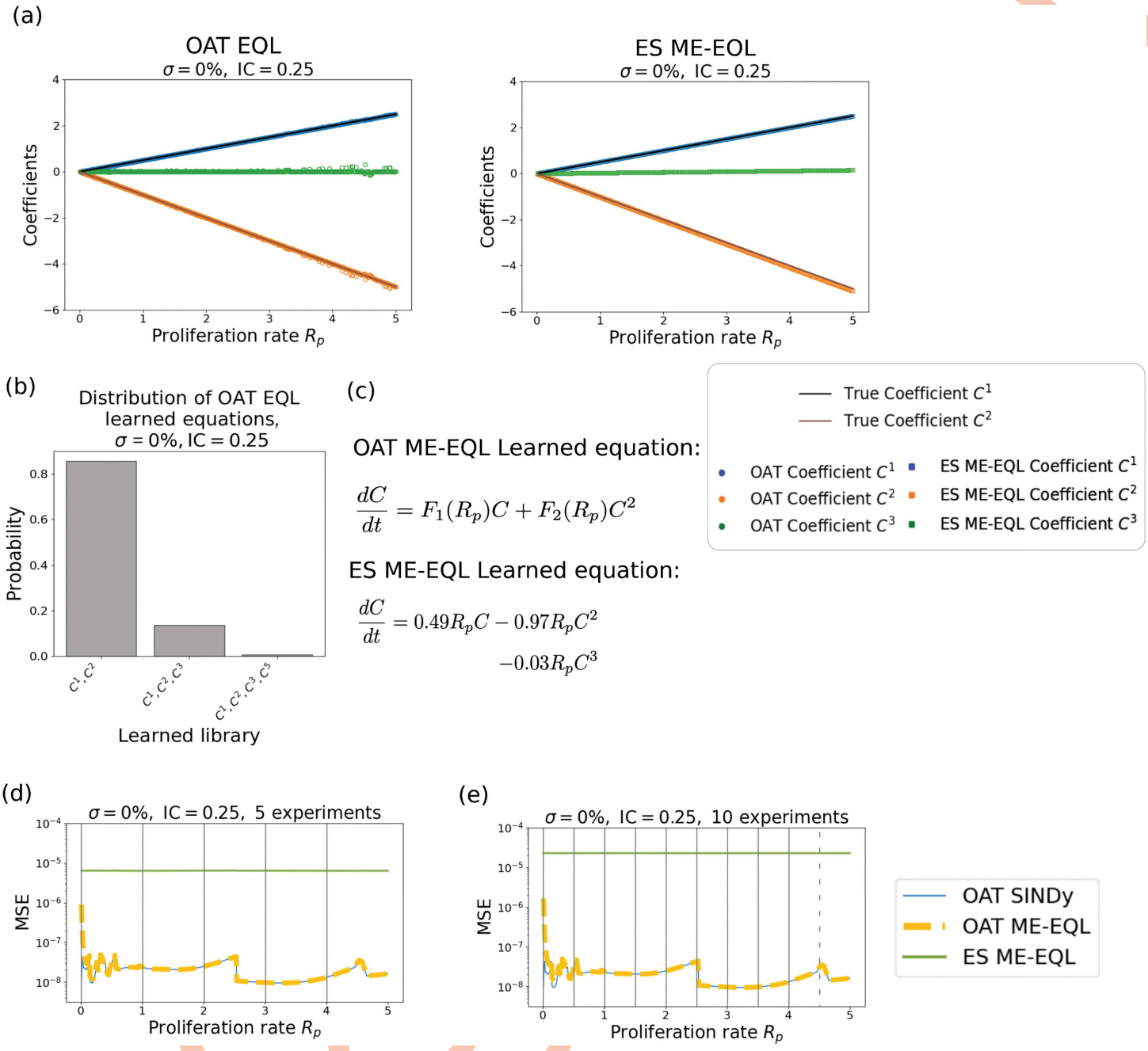

Fig 5. Mean-field learning with IC = 0.25 and σ = 0%. **(a)** Learned Equations using the ES ME-EQL and OAT EQL approaches. **(b)** Most common learned equations from the OAT EQL approach. **(c)** Learned equation from the ES ME-EQL approach. **(d)** MSE of ME-EQL frameworks in predicting mean-field data over all R_p values using 5 experiments. **(e)** MSE of ME-EQL frameworks in predicting mean-field data over all R_p values using 10 experiments.

https://doi.org/10.1371/journal.pcbi.1014161.g005

for two R_p values, as well as the corresponding mean-field DE models (red stars), and the recovered OAT ME-EQL (blue) and ES ME-EQL (green) models. For both R_p values, the mean-field approximation represents a poor match to the ABM data, demonstrating that we no longer have a "ground-truth" model when learning equations for this ABM data, due to the

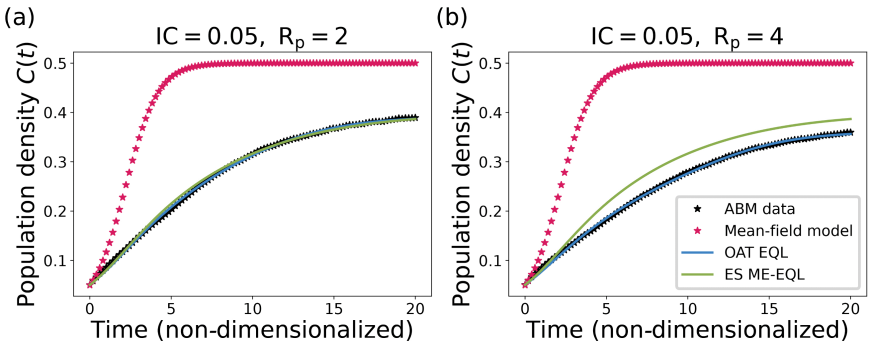

Fig 6. Sample datasets generated using the ABM model (black stars) and fits with the OAT EQL approach (blue line) and the ES ME-EQL approach (green line). The ABM datasets shown here are generated using proliferation rate $R_p = 2$ (a) and $R_p = 4$ (b) and initial condition IC = 0.05.

https://doi.org/10.1371/journal.pcbi.1014161.g006

spatial heterogeneities in the model. We aim to assess whether the two ME-EQL methods can learn ODE models that describe the ABM behavior across parameter space, and to identify parameter regimes where the mean-field model is insufficient to describe the dynamics.

Fig 7 displays the learned coefficients, distribution of learned equations, and final learned equations for the OAT EQL, ES ME-EQL and OAT ME-EQL methods for IC = 0.05 (top) and IC = 0.25 (bottom). For IC = 0.05, the learned coefficients show some continuity with small variation for OAT EQL, and there are clearly differences between the coefficients learned with OAT EQL and ES ME-EQL (Fig 7a). The most commonly-learned OAT EQL equation contains four terms (up to C^5) while the ES ME-EQL equation learns four terms (up to C^4) (Fig 7b,c). There is more noise in the learned coefficients for IC = 0.25, which is due in part to the additional model structures learned with OAT EQL (Fig 7e). The most commonly-learned model structure for OAT EQL is learned for 73% of R_p values (as compared to 92% for IC = 0.05). This model structure contains 3 terms (up to C^4) for OAT ME-EQL, while the ES ME-EQL learned equation recovers a logistic function (Fig 7f).

The ES ME-EQL methodology learns fourth-order and second-order models for initial conditions 0.05 and 0.25, respectively. Interestingly, the model structure learned by ES ME-EQL for the initial condition of 0.05 is not learned for any single parameter set in OAT EQL. Moreover, for IC = 0.25, ES ME-EQL learns the model structure of the mean-field model, but with different parameter values. The parameterized learned models can be found in S2 Table.

We now test the ability of the learned models to predict ABM population data at parameter values not used in learning. Fig 8 displays the results for all models when learning from 10 experiments (i.e., 10 R_p values) and 5 experiments for IC = 0.05. When using 10 R_p values for interpolation, we observe that the interpolated coefficients for OAT ME-EQL are nonlinear (Fig 8a). Of the three ME methods, the OAT ME-EQL learns models with the lowest MSE values for most R_p values, although the mean-field DE model performs best for small R_p values (Fig 8b). Similar results are obtained when using 5 experiments, but we find that the OAT ME-EQL method learns DE models with slightly worse MSE for $R_p < 1$ and $R_p > 4$ (Fig 8c). This indicates the method may be sensitive to the exact R_p values for which data is available.

Fig 9 illustrates the ability of the learned models to generalize to out-of-sample parameters for the ABM data with IC = 0.25. In general, there is less consistency in the learned models over the 10 equally-spaced R_p values considered: only 6 of the 10 experiments share the same model structure in the OAT EQL approach. In particular, the most popular model is consistently learned for larger R_p values ($R_p > 2.5$). As a result, the OAT ME-EQL approach learns predictive models for larger R_p values, but generates worse predictions for smaller R_p values. When learning from 5 experiments, the OAT EQL approach only recovers the most popular model for 2 R_p values. The OAT ME-EQL approach learns a model

UNCORRECTED PROOF

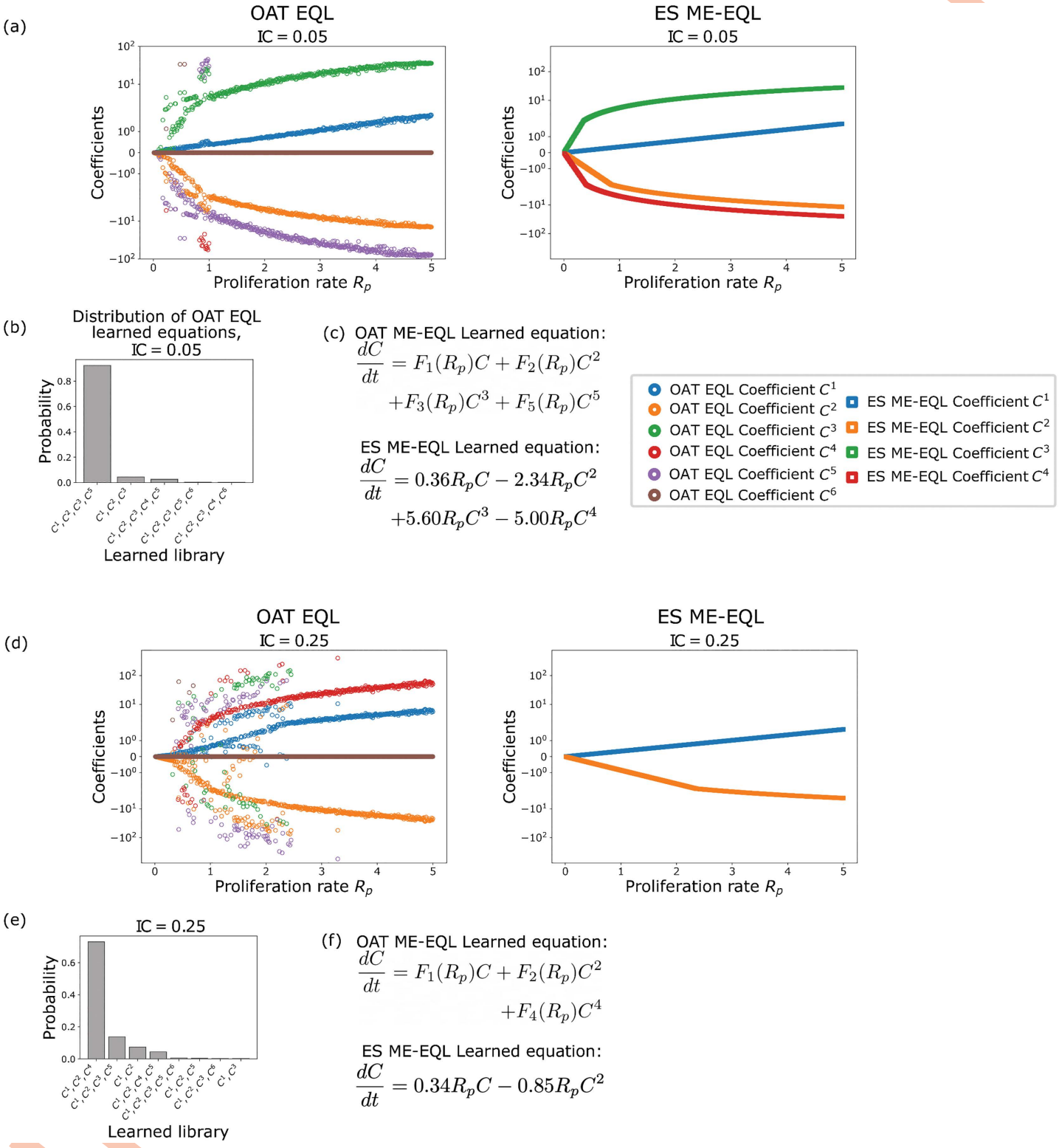

Fig 7. Learned coefficients and models for the agent-based model for IC = 0.05 (top) and IC = 0.25 (bottom). Panels **(a)**, (d) display the learned model coefficients using OAT EQL (colored circles) and learned model coefficients using ES ME-EQL (hollow shapes). Panels **(b)**, (e) depict a histogram of the frequencies of the learned models for OAT EQL. Panels **(c)**, (f) list the learned OAT ME-EQL and ES ME-EQL models. For the IC = 0.05 case **(a-c)**,

there is greater agreement between the OAT EQL and ES ME-EQL learned coefficients, although there are deviations for OAT EQL when $R_p < 0.5$. The most commonly-learned model for OAT EQL (92%) contains four terms (up to C^5), while the ES ME-EQL learned model contains four terms (up to C^4). For the IC = 0.25 case **(d-f)**, there is greater disagreement between the learned coefficients from OAT EQL and ES ME-EQL. There are more deviations from continuity for OAT EQL coefficients and over a larger set of R_p space. The recovered model for ES ME-EQL is logistic, while the most commonly-recovered model for OAT EQL (73%) contains three terms (up to C^4).

https://doi.org/10.1371/journal.pcbi.1014161.g007

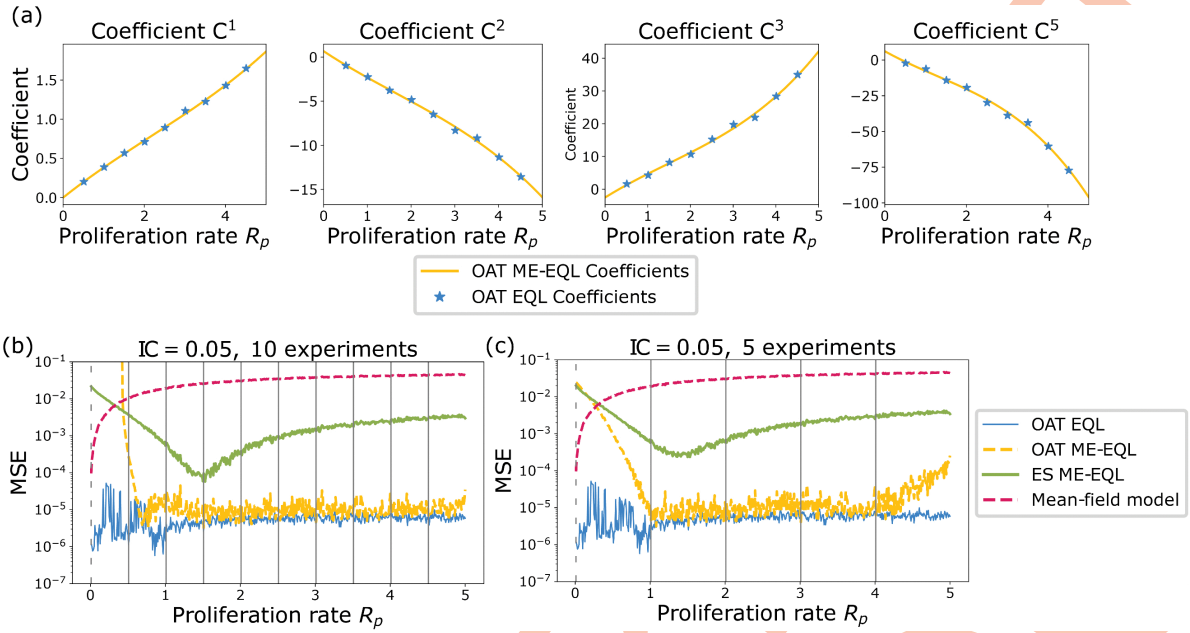

Fig 8. Comparison of the generalizability of learned equations using mean-field model (Eq (2)), OAT ME-EQL, and ES ME-EQL for the ABM Model with IC = 0.05. Panel (a) displays the interpolated coefficients for the OAT ME-EQL method (yellow) using 10 OAT EQL learned parameters (blue stars). Panels **(b)**, (c) display the MSE between data and recovered models for learning from a maximum of 10 R_p values (b) and 5 R_p values **(c)**. The results from the non-generalized OAT EQL for each separate R_p value are shown in blue, the OAT ME-EQL model with interpolated coefficients is shown in yellow dashes, and the ES ME-EQL learning is shown in green solid lines. The mean-field approximation (Eq (2)) is depicted in red dashes. Gray vertical bars indicate the small set of R_p values from which OAT ME-EQL coefficients were learned from. Dashes indicate that OAT ME-EQL did not include the dataset corresponding to that R_p value, since this framework did not learn the most popular model at that parameter value. For very small R_p values, the mean-field approximation results in the lowest MSE for all the generalizable models. However, for all other R_p values, the OAT ME-EQL approach outperforms the mean-field model and ES ME-EQL approaches in generalizability.

https://doi.org/10.1371/journal.pcbi.1014161.g008

whose prediction deteriorates for R_p values outside the range of these two values. Still, the OAT ME-EQL approach outperforms the ES ME-EQL approach in generalizability, with the exception of select small R_p parameter ranges.

Can we infer R_p from ABM data?

Limited sampling in experimental biology, particularly in spatiotemporal cellular and ecological systems, poses significant challenges for using mathematical models to investigate mechanisms generating the observed dynamics. This challenge arises from the need to estimate model parameters from sparse and noisy data [41,42]. Therefore, it is essential to assess whether generalizable surrogate models of ABMs can be learned from a limited number of simulations or experiments. Here, we investigate whether the multi-experiment learning frameworks proposed can yield ODE models that retain predictive accuracy when applied to out-of-sample experimental conditions.

UNCORRECTED PROOF

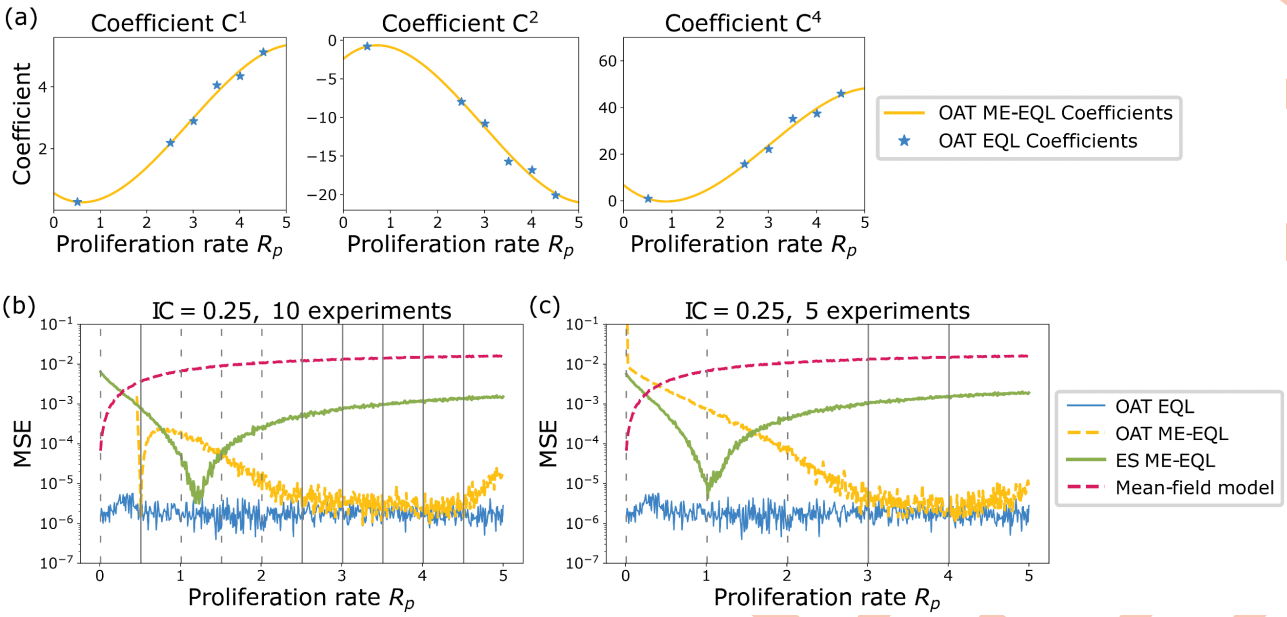

Fig 9. Comparison of the generalizability of learned equations using the mean-field model approximation (Eq (2)), OAT ME-EQL, and ES ME-EQL for the ABM Model with IC = 0.25. Panel (a) displays the interpolated coefficients for the OAT ME-EQL method (yellow) using a maximum of 10 OAT EQL learned parameters (blue stars). Panels **(b)**, (c) display the MSE between data and recovered models for learning from a maximum of 10 R_p values (b) and 5 R_p values **(c)**. The results from the non-generalized OAT EQL for each separate R_p value are shown in blue, the OAT ME-EQL model with interpolated coefficients is shown in yellow dashes, and the ES ME-EQL learning is shown in green solid lines. The mean-field approximation (Eq (2)) is depicted in red dashes. Gray vertical bars indicate the small set of R_p values from which OAT ME-EQL coefficients were learned from. Dashes indicate that OAT ME-EQL did not include the dataset corresponding to that R_p value, since this framework did not learn the most popular model at that parameter value. When examining generalizability, at very small R_p values, the mean-field approximation results in the lowest MSE. However, for all other R_p values, OAT ME-EQL outperforms the mean-field model and ES ME-EQL in generalizability. In contrast to the IC = 0.05 case (Fig 8), there is more variation in learned models for OAT EQL, and thus, the OAT ME-EQL method rejects more learned models (using only 6 out of a maximum of 10) for interpolation.

https://doi.org/10.1371/journal.pcbi.1014161.g009

We investigate the accuracy of each of the three parameterized models in estimating the parameter R_p that generates a single noisy ABM simulation. To assess how this accuracy varies with R_p, we simulated the ABM for 50 different experiments for R_p values between 0.01 and 5.0 for both ICs of 0.05 and 0.25. We estimate the value of R_p that generated a single noisy ABM dataset $C_d(t)$ using a DE model $C(t;R_p)$ by minimizing:

$$\hat{R}_p = \arg\min_{R_p \in \mathbb{R}} \sum_{i=1}^{N} (C_d(t_i) - C(t; R_p))^2. \tag{11}$$

Once we have obtained an estimate $\hat{R}_p$, we estimate its relative error as:

$$\text{Relative } R_p \text{ Error} = \left| \frac{\hat{R}_p - R_p}{R_p} \right|. \tag{12}$$

To understand the uncertainty of our R_p estimate, we calculate 10 separate ABM datasets at each R_p value; we then estimate R_p for each of the 10 noisy datasets. This results in 10 R_p estimates for each R_p value and initial condition, and we report the mean and standard deviation of errors of these values.

UNCORRECTED PROOF

We compare the performance of the mean-field DE model and the learned equations from the OAT ME-EQL and ES ME-EQL pipelines from Learning for agent-based model data that we learned from 10 R_p values (Fig 10 and Table 1). For both initial conditions, the mean-field DE results in the lowest relative error for very small values of R_p, and the two ME-EQL learned model poorly estimate these R_p values. For R_p values above 0.33, however, the mean-field DE obtains higher error values than the two ME-EQL approaches. The OAT ME-EQL learned model achieves the most accurate estimates for most values of R_p above 0.5, although the ES ME-EQL learned model achieves comparable estimates for values of R_p between 1 and 2.

Discussion and conclusions

Agent-based models are a natural means of describing spatiotemporal dynamics in many biological systems, yet the stochastic and parameter-heavy structure of these models presents challenges for inference and analysis. This motivates the development of non-spatial, population-level models to approximate ABMs, whether derived analytically from first principles or through equation learning. In the case of the birth–death–migration dynamics that we considered here, a mean-field model (under the case of well-mixing) is well known, but this ODE does not agree well with ABM simulations in parameter regimes where spatial correlations are significant. At the other extreme, the traditional EQL approach of learning a non-parameterized model for each experiment faces issues related to generalizability and out-of-sample prediction. Motivated by these challenges, we propose two data-driven methods for identifying generalizable and parameterized population-level models from multiple biological experiments. Our intent is to compare and contrast the two methods for this ABM system and understand the distinct challenges associated with each. Our work represents an intermediate framework between deriving a mean-field model from ABM rules and learning a single model for each ABM parameter set.

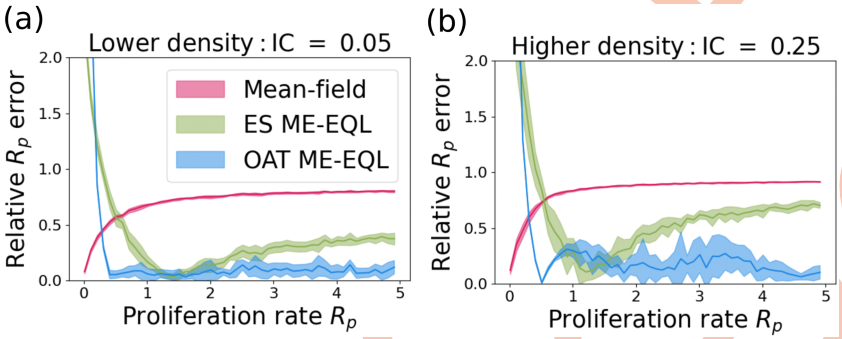

Fig 10. Error in learning ABM parameter R_p from a single out-of-sample ABM simulation using the mean-field DE, ES ME-EQL, and OAT ME-EQL models for (a) IC = 0.05 and (b) IC = 0.25. Mean-field results are shown in magenta, ES ME-EQL is shown in green, and OAT ME-EQL is displayed in blue. Solid bars represent the mean error over 10 ABM simulations, and the shaded area represents one standard deviation about the mean. For small values of R_p, the mean-field model provides the best approximation of R_p. For IC = 0.05, OAT ME-EQL approximations generally perform the best for larger R_p values. For IC = 0.25, there are regions in R_p parameter space for which OAT ME-EQL and ES ME-EQL produce similar errors, but for larger R_p values, OAT ME-EQL produces better fits.

https://doi.org/10.1371/journal.pcbi.1014161.g010

Table 1. Relative R_p error values for various R_p values for all three parameterized models for IC = 0.05. Bold values denote the lowest relative errors (and, in turn, best parameter prediction) for each presented R_p value.

	$R_p = 0.01$	$R_p = 2.51$	$R_p = 4.91$
Mean-field	**0.081**	0.779	0.802
ES ME-EQL	2.273	0.270	0.377
OAT ME-EQL	53.464	**0.142**	**0.121**

https://doi.org/10.1371/journal.pcbi.1014161.t001

UNCORRECTED PROOF

Specifically, to help enhance generalizability in EQL, we considered two ways of learning ODE models from simulations of a birth–death–migration ABM or its associated mean-field model under a range of proliferation rates. Our two methods—OAT and ES ME-EQL—differ in how they incorporate information from multiple experiments. In the OAT ME-EQL approach, we performed equation learning m times for m proliferation rate values, resulting in m models. Post-learning, we interpolated the coefficients for each common learned term as a function of the ABM proliferation rate, resulting in a single model. In our ES approach, we instead embedded the proliferation rate directly in our library and used all m ABM datasets jointly for learning a single equation. Both methods led to parameterized models that closely fit our data. Moreover, in the case of ABM data with strong spatial correlations, our learned models generally outperformed the mean-field model. We evaluated our methods through out-of-sample prediction and found that five experiments (i.e., datasets corresponding to five proliferation rates) were sufficient to learn generalizable models that provide accurate predictions across unobserved proliferation rates.

Population size data are often more readily available than spatial data in complex biological systems, and this makes inferring parameters in a population-level model (as a surrogate for its corresponding ABM) attractive. However, mean-field models are not known for many biological systems, or they may rely on such strong simplifications that insight into parameters at the mean-field level does not translate into insight at the ABM level. OAT and ES ME-EQL help address these challenges and provide a means of estimating agent-based parameters from population data. We found that our methods could be used to recover ABM parameter values by fitting our learned parameterized models to population density in time from noisy ABM simulations. This is a major benefit of learning generalizable models that establish a map between ABM and ODE parameters. Moreover, while a mean-field model is known for the dynamics that we considered, we recovered more accurate ABM proliferation rates using our learned models than using the mean-field model, except in settings with weak spatial structure. This opens up many exciting directions. In the future, it will be interesting to apply our approach to inference in more complicated ABMs for which mean-field models are not known, as well as to estimate rates directly from biological data, similar to the SMoRe ParS framework [11,12]. For example, Gerlee et al. [43] used a spatial ABM of tumor initiation to study how autocrine signaling can generate Allee effects in 2D *in vitro* glioblastoma cell line time series data, while Malik et al. [44] fit reaction–advection–diffusion PDE models to growth curves of multicellular tumor spheroids from patient-derived glioblastoma cell lines to quantify inter-patient and intra-tumor heterogeneity. Our multi-experiment equation learning approach could be applied to such experimental time series to learn differential equations across multiple patient cell lines simultaneously, thereby enhancing the generalizability of the inferred equations governing proliferation, migration, and signaling feedback in glioblastoma cell populations. One could also compare our methods for ABM parameter recovery to inference based on more complicated ODE models that account for spatial correlations in time [17,45].

Overall, our two ME-EQL methods face different challenges. While OAT ME-EQL often resulted in lower mean square errors than did ES ME-EQL, its use of interpolation can be a limitation. In particular, because we interpolate after restricting to the most commonly-learned model structure, OAT ME-EQL may be unreliable when there is high variability in the terms learned. On the other hand, one limitation of ES ME-EQL is its dependence on knowledge of an appropriate library. We included terms with linear dependence on the proliferation rate R_p to match the complexity in our two methods and because birth–death–migration dynamics give rise to a mean-field model with terms that depend linearly on R_p. If there is no information on how model terms depend on the parameters of interest, the library in ES ME-EQL would grow. In this sense, OAT ME-EQL requires less knowledge of the underlying biological dynamics. In the future, it will be interesting to learn from ABM dynamics when a mean-field model exists but our libraries do not include the mean-field model terms. We expect that OAT ME-EQL may be less prone to bias from a poorly chosen library in this setting. Moreover, by enforcing a unified model structure, ES ME-EQL assumes that a single model under varying parameters can explain our biological system. In settings where it is not known if this is the case, OAT ME-EQL could be used first to help identify whether or not a single model applies. After distinguishing regimes of system behavior, we could then apply either OAT ME-EQL or ES ME-EQL to learn a model in each regime.

UNCORRECTED PROOF

Both of our ME-EQL frameworks may benefit from further investigation into their implementation and optimization. We used LASSO for sparse regression and AIC for model selection due to their simplicity and wide usage in the EQL literature [25,36]. Other options for sparse regression include ridge regression [46], greedy methods [47], or sparse relaxed regularized regression (SR3) [48]; see [49] for a comprehensive review. While LASSO is intended to perform sparse regression, it may lead to parameter shrinkage without thresholding small parameters to zero [50]. To combat this, methods such as sequentially thresholded least squares (STLSQ) perform multiple iterations of ridge regression and parameter thresholding [51–53], either based on parameter values themselves or the magnitude of the associated terms [52,54]. In contrast to LASSO, which requires the selection of a single hyperparameter for regularization strength, methods like STSQL require additional hyperparameters to implement thresholding. In future work, methods such as Bayesian optimization [55] could be explored to overcome the computational bottleneck associated with selecting hyperparameters in higher dimensions. In our ME-EQL frameworks, we also used domain knowledge to avoid large coefficient values in the learned models that would be unrealistic for the system studied here. In future work, it would be interesting to determine how the results change with no prior knowledge, and whether this impacts the comparison of the learned models from OAT and ES ME-EQL.

There are many ways to build on our study, and we highlight several here. For example, it will be interesting to further consider the role of noise in EQL. By first learning from mean-field model data with varying noise levels, we determined how our two methods perform when there is a known (e.g., *correct*) model to learn. Surprisingly, we found that, even in the case of noise-free data, single-experiment EQL and our two ME-EQL methods did not always learn the terms in the correct underlying model. In the future, it would be interesting to investigate the robustness of ME-EQL to the number of ABM simulations, as increasing the number of simulations would result in smoother data but be more computationally expensive. Learning from fewer ABM runs would also be of interest, given the typical noise levels in available experimental data. Our successful inference of parameter R_p from a single, noisy, and out-of-sample ABM simulation in Fig 10 suggests that learning from fewer datasets could also yield ODE models that are generalizable and predictive. We could also consider improved methods for numerically calculating derivatives from noisy data. While we chose to base our work on SINDy [22], related methods such as weak SINDy [54,56] could be used to overcome challenges related to approximating the time derivatives needed for equation learning. Incorporating weak SINDy into our ES ME-EQL approach would be a potential avenue for future work to improve upon the results presented here.

Another natural extension of our approach would be to consider multiple independently varying parameters. In this setting, OAT ME-EQL could scale relatively well, since models could still be learned separately for each parameter combination, and coefficients interpolated to predict dynamics for unobserved parameter sets. ES ME-EQL could also be extended, but including multiple parameters in the candidate library would increase its size combinatorially, raising computational costs and reducing interpretability.

Finally, another exciting direction for future work is refining our choice of ABM parameter values to determine the most informative experiments for ME-EQL. This would inform experimental design, but is also related to sensitivity analysis, which could be used to show that our learned model is more sensitive at low proliferation rates and suggest that more experiments should be performed there. More broadly, the major benefit of learning generalizable, parameterized models is that they are amenable to traditional, powerful approaches such as bifurcation analysis, optimal control, and uncertainty quantification. We expect that combining equation learning with such classic modeling approaches will shed new light on biological systems in the future.

Supporting information

S1 Text. Calculating the optimal regularization hyperparameter λ. Details on methods to calculate the optimal regularization hyperparameter λ and the procedure used to select the final model.
(PDF)

UNCORRECTED PROOF

S2 Text. Algorithmic differences between OAT ME-EQL and ES ME-EQL. Algorithms for the OAT ME-EQL and ES ME-EQL approaches.
(PDF)

S1 Fig. Example plots of the hyperparameter $\bar{\lambda}$ plotted against the AIC scores of $R_p = 1$ (left) and $R_p = 5$ (right). Optimal λ selected in red square. Text indicates the learned model structure at each jump in the plot.
(PDF)

S2 Fig. Example model simulations highlighting how initial conditions affect information content. Snapshots of ABM simulations of birth, death, and migration dynamics at different timepoints for initial conditions with 5% and 25% of sites occupied, respectively, and corresponding mean population sizes in time across 25 ABM simulations.
(PDF)

S1 Table. Models learned using ME-EQL for mean-field model data. Models learned using ME-EQL methods for the mean-field model data with initial conditions 0.05 and 0.25, and for noise levels σ = 0% and σ = 0.25%.
(PDF)

S2 Table. Models learned using ME-EQL for ABM data. Models learned using ME-EQL methods for the ABM data with initial conditions 0.05 and 0.25.
(PDF)

Acknowledgments

This project began during a SQuaRE at the American Institute for Mathematics. The authors thank AIM for providing a supportive and mathematically rich environment. The authors also acknowledge use of the ELSA high performance computing cluster at The College of New Jersey for conducting the research reported in this paper.

Author contributions

Conceptualization: Maria-Veronica Ciocanel, John T. Nardini, Kevin B. Flores, Erica M. Rutter, Suzanne S. Sindi, Alexandria Volkening.

Methodology: Maria-Veronica Ciocanel, John T. Nardini, Kevin B. Flores, Erica M. Rutter, Suzanne S. Sindi, Alexandria Volkening.

Software: Maria-Veronica Ciocanel, John T. Nardini.

Validation: Maria-Veronica Ciocanel, John T. Nardini.

Writing – original draft: Maria-Veronica Ciocanel, John T. Nardini, Kevin B. Flores, Erica M. Rutter, Suzanne S. Sindi, Alexandria Volkening.

Writing – review & editing: Maria-Veronica Ciocanel, John T. Nardini, Kevin B. Flores, Erica M. Rutter, Suzanne S. Sindi, Alexandria Volkening.

References

1. Wadkin LE, Orozco-Fuentes S, Neganova I, Lako M, Shukurov A, Parker NG. The recent advances in the mathematical modelling of human pluripotent stem cells. SN Appl Sci. 2020;2(2):276. https://doi.org/10.1007/s42452-020-2070-3 PMID: 32803125
2. Styles KM, Brown AT, Sagona AP. A Review of Using Mathematical Modeling to Improve Our Understanding of Bacteriophage, Bacteria, and Eukaryotic Interactions. Front Microbiol. 2021;12:724767. https://doi.org/10.3389/fmicb.2021.724767 PMID: 34621252
3. Rangamani P, Iyengar R. Modelling spatio-temporal interactions within the cell. J Biosci. 2007;32:157–67. https://doi.org/10.1007/s12038-007-0014-3

UNCORRECTED PROOF

4. Volkening A. Linking genotype, cell behavior, and phenotype: multidisciplinary perspectives with a basis in zebrafish patterns. Curr Opin Genet Dev. 2020;63:78–85. https://doi.org/10.1016/j.gde.2020.05.010 PMID: 32604031
5. Edelstein-Keshet L. Mathematical models in biology. Philadelphia, PA: SIAM; 2005.
6. Murray JD. Mathematical biology: I. An introduction. New York: Springer Science & Business Media; 2007.
7. Friedman A, Kao CY. Mathematical Modeling of Biological Processes. vol. 1 of Lecture Notes on Mathematical Modelling in the Life Sciences. Cham: Springer; 2014.
8. Norton KA, Bergman D, Jain HV, Jackson T. Advances in Surrogate Modeling for Biological Agent-Based Simulations: Trends, Challenges, and Future Prospects. arXiv preprint arXiv:250411617. 2025.
9. Fonseca LL, Böttcher L, Mehrad B, Laubenbacher RC. Optimal control of agent-based models via surrogate modeling. PLoS Comput Biol. 2025;21(1):e1012138. https://doi.org/10.1371/journal.pcbi.1012138 PMID: 39808665
10. Nardini JT. Forecasting and Predicting Stochastic Agent-Based Model Data with Biologically-Informed Neural Networks. Bull Math Biol. 2024;86(11):130. https://doi.org/10.1007/s11538-024-01357-2 PMID: 39307859
11. Jain HV, Norton K-A, Prado BB, Jackson TL. SMoRe ParS: A novel methodology for bridging modeling modalities and experimental data applied to 3D vascular tumor growth. Front Mol Biosci. 2022;9:1056461. https://doi.org/10.3389/fmolb.2022.1056461 PMID: 36619168
12. Bergman DR, Norton K-A, Jain HV, Jackson T. Connecting Agent-Based Models with High-Dimensional Parameter Spaces to Multidimensional Data Using SMoRe ParS: A Surrogate Modeling Approach. Bull Math Biol. 2023;86(1):11. https://doi.org/10.1007/s11538-023-01240-6 PMID: 38159216
13. Kiss IZ, Miller JC, Simon PL. Mathematics of epidemics on networks: from exact to approximate models. vol. 46 of Interdisciplinary Applied Mathematics. Cham: Springer; 2017.
14. Plank MJ, Law R. Spatial point processes and moment dynamics in the life sciences: a parsimonious derivation and some extensions. Bull Math Biol. 2015;77(4):586–613. https://doi.org/10.1007/s11538-014-0018-8 PMID: 25216969
15. Johnston ST, Simpson MJ, Baker RE. Mean-field descriptions of collective migration with strong adhesion. Phys Rev E Stat Nonlin Soft Matter Phys. 2012;85(5 Pt 1):051922. https://doi.org/10.1103/PhysRevE.85.051922 PMID: 23004802
16. Erban R, Chapman SJ. Stochastic Modelling of Reaction–Diffusion Processes. Cambridge Texts in Applied Mathematics. Cambridge: Cambridge University Press; 2020.
17. Baker RE, Simpson MJ. Correcting mean-field approximations for birth-death-movement processes. Phys Rev E Stat Nonlin Soft Matter Phys. 2010;82(4 Pt 1):041905. https://doi.org/10.1103/PhysRevE.82.041905 PMID: 21230311
18. Nardini JT, Baker RE, Simpson MJ, Flores KB. Learning differential equation models from stochastic agent-based model simulations. J R Soc Interface. 2021;18(176):20200987. https://doi.org/10.1098/rsif.2020.0987 PMID: 33726540
19. Gaskin T, Pavliotis GA, Girolami M. Neural parameter calibration for large-scale multiagent models. Proc Natl Acad Sci U S A. 2023;120(7):e2216415120. https://doi.org/10.1073/pnas.2216415120 PMID: 36763529
20. Koza JR. Genetic programming as a means for programming computers by natural selection. Stat Comput. 1994;4(2). https://doi.org/10.1007/bf00175355
21. Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. Science. 2009;324(5923):81–5.
22. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc Natl Acad Sci U S A. 2016;113(15):3932–7. https://doi.org/10.1073/pnas.1517384113 PMID: 27035946
23. Udrescu S-M, Tegmark M. AI Feynman: A physics-inspired method for symbolic regression. Sci Adv. 2020;6(16):eaay2631. https://doi.org/10.1126/sciadv.aay2631 PMID: 32426452
24. Makke N, Chawla S. Interpretable scientific discovery with symbolic regression: a review. Artif Intell Rev. 2024;57(1). https://doi.org/10.1007/s10462-023-10622-0
25. Mangan NM, Brunton SL, Proctor JL, Kutz JN. Inferring Biological Networks by Sparse Identification of Nonlinear Dynamics. IEEE Trans Mol Biol Multi-Scale Commun. 2016;2(1):52–63. https://doi.org/10.1109/tmbmc.2016.2633265
26. Prokop B, Gelens L. From biological data to oscillator models using SINDy. iScience. 2024;27(4):109316. https://doi.org/10.1016/j.isci.2024.109316 PMID: 38523784
27. Lu L, Jin P, Pang G, Zhang Z, Karniadakis GE. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. Nat Mach Intell. 2021;3(3):218–29. https://doi.org/10.1038/s42256-021-00302-5
28. Johnston ST, Simpson MJ, McElwain DLS. How much information can be obtained from tracking the position of the leading edge in a scratch assay? J R Soc Interface. 2014;11(97):20140325. https://doi.org/10.1098/rsif.2014.0325 PMID: 24850906
29. Bernoff AJ, Topaz CM. Nonlocal Aggregation Models: A Primer of Swarm Equilibria. SIAM Rev. 2013;55(4):709–47. https://doi.org/10.1137/130925669
30. FISHER RA. The wave of advance of advantageous genes. Ann Eugen. 1937;7(4):355–69. https://doi.org/10.1111/j.1469-1809.1937.tb02153.x
31. Swanson KR, Bridge C, Murray JD, Alvord Jr EC. Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion. J Neurol Sci. 2003;216(1):1–10. https://doi.org/10.1016/j.jns.2003.06.001

UNCORRECTED PROOF

32. Simpson MJ, McCue SW. Fisher–KPP-type models of biological invasion: open source computational tools, key concepts and analysis. Proc R Soc A. 2024;480(2294). https://doi.org/10.1098/rspa.2024.0186
33. Nicolaou ZG, Huo G, Chen Y, Brunton SL, Kutz JN. Data-driven discovery and extrapolation of parameterized pattern-forming dynamics. Phys Rev Res. 2023;5(4). https://doi.org/10.1103/physrevresearch.5.l042017
34. Liu Y, Suh K, Maini PK, Cohen DJ, Baker RE. Parameter identifiability and model selection for partial differential equation models of cell invasion. J R Soc Interface. 2024;21(212):20230607. https://doi.org/10.1098/rsif.2023.0607 PMID: 38442862
35. Lagergren JH, Nardini JT, Michael Lavigne G, Rutter EM, Flores KB. Learning partial differential equations for biological transport models from noisy spatio-temporal data. Proc Math Phys Eng Sci. 2020;476(2234):20190800. https://doi.org/10.1098/rspa.2019.0800 PMID: 32201481
36. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Stat Methodol. 1996;58(1):267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
37. Mangan NM, Kutz JN, Brunton SL, Proctor JL. Model selection for dynamical systems via sparse regression and information criteria. Proc Math Phys Eng Sci. 2017;473(2204):20170009. https://doi.org/10.1098/rspa.2017.0009 PMID: 28878554
38. de Silva B, Champion K, Quade M, Loiseau J-C, Kutz J, Brunton S. PySINDy: A Python package for the sparse identification of nonlinear dynamical systems from data. JOSS. 2020;5(49):2104. https://doi.org/10.21105/joss.02104
39. Kaptanoglu A, de Silva B, Fasel U, Kaheman K, Goldschmidt A, Callaham J, et al. PySINDy: A comprehensive Python package for robust sparse system identification. JOSS. 2022;7(69):3994. https://doi.org/10.21105/joss.03994
40. Fasel U, Kutz JN, Brunton BW, Brunton SL. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. Proc Math Phys Eng Sci. 2022;478(2260):20210904. https://doi.org/10.1098/rspa.2021.0904 PMID: 35450025
41. Harrison JU, Baker RE. The impact of temporal sampling resolution on parameter inference for biological transport models. PLoS Comput Biol. 2018;14(6):e1006235. https://doi.org/10.1371/journal.pcbi.1006235 PMID: 29939995
42. Lagergren JH, Nardini JT, Baker RE, Simpson MJ, Flores KB. Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. PLoS Comput Biol. 2020;16(12):e1008462. https://doi.org/10.1371/journal.pcbi.1008462 PMID: 33259472
43. Gerlee P, Altrock PM, Malik A, Krona C, Nelander S. Autocrine signaling can explain the emergence of Allee effects in cancer cell populations. PLoS Comput Biol. 2022;18(3):e1009844. https://doi.org/10.1371/journal.pcbi.1009844 PMID: 35239640
44. Malik AA, Nguyen KC, Nardini JT, Krona CC, Flores KB, Nelander S. Mathematical modeling of multicellular tumor spheroids quantifies inter-patient and intra-tumor heterogeneity. NPJ Syst Biol Appl. 2025;11(1):20. https://doi.org/10.1038/s41540-025-00492-3 PMID: 39955270
45. Nardini JT, Lagergren JH, Hawkins-Daarud A, Curtin L, Morris B, Rutter EM, et al. Learning Equations from Biological Data with Limited Time Samples. Bull Math Biol. 2020;82(9):119. https://doi.org/10.1007/s11538-020-00794-z PMID: 32909137
46. Marquardt DW, Snee RD. Ridge regression in practice. Am Statist. 1975;29(1):3–20. https://doi.org/10.1080/00031305.1975.10479105
47. Zhang T. Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations. IEEE Trans Inform Theory. 2011;57(7):4689–708. https://doi.org/10.1109/tit.2011.2146690
48. Zheng P, Askham T, Brunton SL, Kutz JN, Aravkin AY. A Unified Framework for Sparse Relaxed Regularized Regression: SR3. IEEE Access. 2019;7:1404–23. https://doi.org/10.1109/access.2018.2886528
49. Bertsimas D, Pauphilet J, Van Parys B. Sparse regression: scalable algorithms and empirical performance. Stat Sci. 2020;35(4). https://doi.org/10.1214/19-STS701
50. van de Geer S, Bühlmann P, Zhou S. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). Electron J Statist. 2011;5(none). https://doi.org/10.1214/11-ejs624
51. Rudy SH, Brunton SL, Proctor JL, Kutz JN. Data-driven discovery of partial differential equations. Sci Adv. 2017;3(4):e1602614. https://doi.org/10.1126/sciadv.1602614 PMID: 28508044
52. Kang SH, Liao W, Liu Y. IDENT: Identifying Differential Equations with Numerical Time evolution. 2019. Available from: https://arxiv.org/abs/1904.03538
53. Schneider T, Stuart AM, Wu J-L. Ensemble Kalman inversion for sparse learning of dynamical systems from time-averaged data. Journal of Computational Physics. 2022;470:111559. https://doi.org/10.1016/j.jcp.2022.111559
54. Messenger DA, Bortz DM. Weak SINDy: Galerkin-based Data-driven Model Selection. Multiscale Model Simul. 2021;19(3):1474–97. https://doi.org/10.1137/20m1343166 PMID: 38239761
55. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25; 2012. p. 2951–59.
56. Messenger DA, Bortz DM. Weak sindy for partial differential equations. J Comput Phys. 2021;443:110525. https://doi.org/10.1016/j.jcp.2021.110525 PMID: 34744183

UNCORRECTED PROOF