Data driven model discovery and interpretation for CAR T-cell killing using sparse identification and latent variables

Alexander B. Brummer [1,2,*], **Agata Xella** [3], **Ryan Woodall** [1], **Vikram Adhikarla** [1], **Heyrim Cho** [4], **Margarita Gutova** [5], **Christine E. Brown** [3,*], **Russell C. Rockne** [1,*]

[1] *Division of Mathematical Oncology, Department of Computational and Quantitative Medicine, Beckman Research Institute, City of Hope National Medical Center, Duarte, CA 91010, USA*

[2] *Department of Physics and Astronomy, College of Charleston, Charleston, SC 29401, USA*

[3] *Department of Hemtaology and Hematopoietic Cell Translation and Immuno-Oncology, Beckman Resarch Institute, City of Hope National Medical Center, Duarte, CA 91010, USA*

[4] *Department of Mathematics, University of California, Riverside, CA 92521, USA*

[5] *Department of Stem Cell Biology and Regenerative Medicine, Beckman Research Institute, City of Hope National Medical Center, Duarte, CA 91010, USA*

Correspondence*:
Alexander B. Brummer
Department of Physics and Astronomy, College of Charleston, 58 Coming St, Charleston, SC, USA
brummerab@cofc.edu

Christine E. Brown
Department of Hemtaology and Hematopoietic Cell Translation and Immuno-Oncology, Beckman Resarch Institute, City of Hope National Medical Center, 1500 E Duarte Rd, Duarte, CA 91010, USA, cbrown@coh.org

Russell C. Rockne
Division of Mathematical Oncology, Department of Computational and Quantitative Medicine, Beckman Research Institute, City of Hope National Medical Center, 1500 E Duarte Rd, Duarte, CA 91010, USA, rrockne@coh.org

ABSTRACT

In the development of cell-based cancer therapies, quantitative mathematical models of cellular interactions are instrumental in understanding treatment efficacy. Efforts to validate and interpret mathematical models of cancer cell growth and death hinge first on proposing a precise mathematical model, then analyzing experimental data in the context of the chosen model. In this work, we present the first application of the sparse identification of non-linear dynamics (SINDy) algorithm to a real biological system in order discover cell-cell interaction dynamics in *in vitro* experimental data, using chimeric antigen receptor (CAR) T-cells and patient-derived glioblastoma cells. By combining the techniques of latent variable analysis and SINDy, we infer key aspects of the interaction dynamics of CAR T-cell populations and cancer. Importantly, we show how the model terms can be interpreted biologically in relation to different CAR T-cell functional responses, single or double CAR T-cell-cancer cell binding models, and density-dependent growth dynamics in either of the CAR T-cell or cancer cell populations. We show

how this data-driven model-discovery based approach provides unique insight into CAR T-cell dynamics when compared to an established model-first approach. These results demonstrate the potential for SINDy to improve the implementation and efficacy of CAR T-cell therapy in the clinic through an improved understanding of CAR T-cell dynamics.

Keywords: dynamical systems; latent variables; CAR T-cells; antigen binding; allee effect; SINDy; glioblastoma; cell therapy

1 INTRODUCTION

Dynamical systems modeling is one of the most successfully implemented methodologies throughout mathematical oncology (1). Applications of these *model first* approaches have lead to important insights in fundamental cancer biology as well as the planning and tracking of treatment response for patient cohorts (2, 3, 4, 5, 6, 7, 8, 9). Simultaneously, the last twenty years have seen explosive growth in the study and application of data-driven methods. These *data first* approaches, initially implemented as machine learning methods for imaging and genomics analyses, have seen much success (10, 11). However, such approaches are often limited to classification problems and fall short when the intention is to identify and validate mathematical models of the underlying dynamics. Recent efforts by us and others have aimed to develop methodologies that bridge these *model first* and *data first* approaches. (12, 13, 14).

In this work, we combine the methods of latent variable discovery and sparse identification of nonlinear dynamics (SINDy) (15, 16, 17) to analyze experimental *in vitro* cell killing assay data for chimeric antigen receptor (CAR) T-cells and glioblastoma cancer cells (18). This experimental data, featuring high temporal resolution, offers a unique opportunity to conduct an *in situ* test of the SINDy model discovery method. Interpretation of the discovered SINDy model is conducted under the expectation of a predator-prey interaction model in which the cancer cells function as the prey and the CAR T-cells the predator (19).

Predator-prey systems are a broad class of ordinary differential equations (ODEs) that aim to characterize changes in populations between two or more groups of organisms in which at least one survives via predation on another. Originally applied to the study of plant herbivory (20) and fishery monitoring (21) in the early 20^{th} century, predator-prey models have since become a workhorse of ecology, evolutionary biology, and most recently mathematical oncology (19, 22). For example, predator-prey models of CAR T-cell killing dynamics have shed light on the underlying biological mechanisms of action (18, 7), and have informed effective dosing strategies for combination CAR T-cell and targeted radionuclide therapy (23), and CAR T-cell therapy in combination with the anti-inflammatory steroid Dexamethasone (24). Over time, important extensions to predator-prey models have been incorporated to account for a variety of biological phenomena. In this work we focus on the following extensions: predator growth that is dependent on the density of prey, also known as a functional response (25, 26); individual predator and prey growth that saturates at some maximum value (logistic growth) (18), or has a population threshold below which collapse occurs (the Allee effect) (27, 28); and predator-prey interactions in which one or two CAR T-cells are bound to a single cancer cell at once, referred to as single or double binding, respectively (29, 30).

An ever-present challenge to quantitative biologists is fitting a proposed model to experimental data, also known as parameter estimation or model inference. On one hand, quantitative biologists seek models that capture as much biological realism and complexity as possible.

On the other hand, increasing model complexity increases the computational challenge to accurately, confidently, and expediently determine model parameter values. This approach is further complicated if a researcher chooses to compare competing or complementary models (31, 32). An alternative approach, examined in this paper, is to leverage newly developed methods rooted in data science and machine learning which identify the strength of individual mathematical terms as candidates for an explanatory model. These methods are often referred to as dynamic mode decomposition, symbolic regression, or sparse identification.

Dynamic mode decomposition (DMD) is a data driven technique that interrogates time-series data by performing a singular value decomposition (SVD) on carefully structured matrices of the given data (33, 13). In this formalism, the orthonormal basis vectors generated by singular value decomposition serve as linear generators of the system dynamics such that forward prediction can be performed absent a known underlying mathematical model. Alternatively, SINDy identifies the specific mathematical terms that give rise to the observed dynamics governed by ordinary and partial differential equation models (15). SINDy achieves this by regressing experimental data onto a high-dimensional library of candidate model terms, and it has proven successful in climate modeling (34), fluid mechanics (35), and control theory (36). Since the initial publication of SINDy, several extensions have been studied, including: discovery of rational ordinary differential equations (37, 38); robust implementation with under-sampled data (39) or excessive noise (40); or incorporation of physics informed neural networks when particular symmetries are known to exist (41).

In this paper we utilize our experimental data to test several aspects of the DMD and SINDy frameworks. In Section 2.2 we introduce the families of models that are anticipated to be simultaneously biologically relevant and identifiable by SINDy, and we introduce a new approach to performing SINDy-based model inference.

In Section 2.3.1 we present the latent variable analysis based on DMD that is used to generate the time-series CAR T-cell trajectories based on those of the cancer cells and the known boundary values for the CAR T-cells. In Section 2.3.2 we introduce the SINDy methodology in the particular context of our application. Results of our approach are presented in Section 3 where we (1) highlight how the discovered models vary as a result of different initial conditions in the cancer cell and CAR T-cell populations and (2) examine how well the discovered models found in this *data first* approach compare to a typical *model first* in characterizing the experimental data. In Section 4 we demonstrate how our results can guide experimental design to validate the predictions made by the discovered models, and we elaborate on some of the challenges encountered in this study.

2 MATERIALS AND METHODS

2.1 Experimental setup

The data analyzed in this study come from previously conducted experiments whose procedures are described in Sahoo et al. (18) and Brummer et al. (24), and summarized in Figure 1. The primary brain tumor cell line studied (PBT128) was selected for its endogenous high and relatively uniform expression of IL13Rα2 antigen (89.11% IL13Rα2+) (24). This cell line was derived from glioblastoma tumor resection tissue as described in (42, 43). To generate IL13Rα2-targeted CAR T-cell lines, healthy donor CD62L+ naive and memory T-cells were lentivirally transduced to

express second-generation 4-1BB-containing CAR that utilizes the IL13 cytokine with an E12Y engineered mutation as the IL13Rα2 targeting domain (44).

Cell killing experiments were conducted and monitored with an xCELLigence cell analyzer system. Measurements of cancer cell populations are reported every 15 minutes through changes in electrical impedance as cancer cells adhere to microelectrode plates, and are reported in units of Cell Index (CI), where 1 CI $\approx 10K$ cells (45, 46, 47). Flow cytometry was used to count the non-adherent CAR T-cells upon termination of the experiment. Measurements of CAR T-cell populations are reported in units of CI for the purposes of working in a common scale. We used the conversion factor of 1 CI $\approx 10K$ cells. Cancer cells were seeded at $10K - 20K$ cells and left either untreated or treated with only CAR T-cells, with treatments occurring 24 hours after seeding and monitored for 6-8 days (Figure 1). CAR T-cell treatments were performed with effector-to-target ratios (E:T) of 1:4, 1:8, and 1:20. All conditions were conducted in duplicate.

2.2 Effective interaction models

Challenges to the *model first* approach to systems biology are (1) deciding on a sufficiently comprehensive model that captures all pertinent phenomena and (2) fitting the selected model to available data. Researchers are tasked with justifying their decisions in selecting candidate models. Yet, a common feature of dynamical systems models are the presence of ratios of polynomials. Such terms in ODEs can be difficult for the convergence of optimization algorithms to global solutions due to the possible existence of multiple local solutions within the model parameter space (48). In such instances researchers must either rely on high performance computational methods, have collected a vast amount of experimental data, or both. To address this problem, we utilize binomial expansions of candidate model terms under the assumptions of CAR T-cell treatment success and fast, irreversible reaction kinetics. In the following sections we present the space of possible models anticipated to characterize our experimental system, and the steps necessary to reduce the complexity of these candidate models.

The dynamical model that our experimental system is anticipated to follow is defined generically as,

$$\frac{dx}{dt} = \mathcal{G}_x(x) - x\mathcal{B}_x(y) \tag{1}$$

$$\frac{dy}{dt} = \mathcal{G}_y(y) + y\mathcal{R}(x) - x\mathcal{B}_y(y) \tag{2}$$

where $\mathcal{G}_x$ and $\mathcal{G}_y$ represent a growth-death model for the cancer cells, x, and the CAR T-cells, y. $\mathcal{B}_x$ and $\mathcal{B}_y$ represent a binding model for whether single or pairs of CAR T-cells attack individual cancer cells, and $\mathcal{R}$ represents a model for the CAR T-cell functional response. In the subsections below, we explore different families of models representing the terms in the above equations. Explicitly, we examine different types of (a) Growth and death models, (b) Functional response models, and (c) CAR T-cell-cancer cell binding models.

2.2.1 Growth and Death

We consider three different growth-death models for both the cancer cells and CAR T-cells. These are logistic growth, and the weak and strong Allee effect models, presented as,

$$\mathcal{G}_x(x) = \begin{cases} \rho_x x\left(1-\dfrac{x}{K_x}\right) & \text{Logistic growth} \quad (3) \\ \rho_x x\left(1-\dfrac{x}{K_x}\right)\left(1+\dfrac{x}{A_x}\right) & \text{Weak Allee effect} \quad (4) \\ \rho_x x\left(1-\dfrac{x}{K_x}\right)\left(\dfrac{x}{B_x}-1\right) & \text{Strong Allee effect} \quad (5) \end{cases}$$

for $\mathcal{G}_x(x)$, and similarly for $\mathcal{G}_y(y)$. Here, ρ_x is the net growth rate, K_x is the population carrying capacity, A_x is a weak parameterization of deviations from logistic growth, and B_x is the threshold for population survival or death absent predation. All model parameters are assumed positive, with the added constraint that $K_x > B_x > 0$. We anticipate similar growth models for the CAR T-cells, $\mathcal{G}_y(y)$, with allowance of different models for the different cell types and model constants. Logistic growth is commonly favored for its simplicity in experimental systems (18, 23, 24), while there is growing evidence that Allee effects are required for accurate characterization of low density cancer cell populations (27, 28, 49, 50) or as the result of directed movement (51), the latter of which being an observable feature of CAR T-cell behavior using bright field imaging (18, 24).

In Figure 2, graphs of population growth rates versus population size and population size versus time are presented for each growth model and for a variety of initial conditions. Parameter values used were $\rho = 0.75$ hrs^{-1}, $K = 10$ CI, $A = 5$ CI, and $B = 5$ CI. Examination of the logistic growth model in Figure 2**a**, **b** and the weak Allee effect in Figure 2**c**, **d** demonstrates similar population saturation at the carrying capacity $K = 10$ CI, but a slight deviation between how the models reach saturation. Specifically, the weak Allee effect exhibits a reduced per capita growth rate at low population densities comapared to logistic growth. Examination of Figure 2**e** and **d** demonstrates the crucial difference between the strong Allee effect and either of the logistic growth or weak Allee effect through the existence of a minimum population threshold, B, above which the population will persist, and below which the population will die off.

Due to the fact that SINDy produces discovered models in their polynomial form without factoring, or grouping of terms together, we must consider the un-factored polynomial form of each model. To determine appropriate constraints on the model coefficients, we will expand the growth models and factor by common monomials. Doing so for $\mathcal{G}_x(x)$ and dropping the subscript gives the following,

$$\mathcal{G}_x(x) = \begin{cases} \rho_x x - \dfrac{\rho_x}{K_x} x^2 & \text{Logistic growth} \quad (6) \\ \rho_x x + \left(\dfrac{\rho_x}{A_x} - \dfrac{\rho_x}{K_x} \right) x^2 - \dfrac{\rho_x}{K_x A_x} x^3 & \text{Weak Allee effect} \quad (7) \\ -\rho_x x + \left(\dfrac{\rho_x}{K_x} + \dfrac{\rho_x}{B_x} \right) x^2 - \dfrac{\rho_x}{K_x B_x} x^3 & \text{Strong Allee effect} \quad (8) \end{cases}$$

and similarly for $\mathcal{G}_y(y)$. Here we can see that the coefficients for x and x^2 can be positive or negative, but the coefficients for x^3 must be fixed as negative values, where we have absorbed the minus signs in Eqs. (7)-(8) into $\rho_x / K_x A_x$.

2.2.2 Functional response

We next consider the first three types of functional response models that characterize how the CAR T-cells respond, or expand, in the presence of cancer cells. These models are defined as,

$$\mathcal{R}(x) = \begin{cases} px & \text{Type I} \quad (9) \\ \dfrac{px}{g + x} & \text{Type II} \quad (10) \\ \dfrac{px^2}{g^2 + x^2} & \text{Type III} \quad (11) \end{cases}$$

where p is the predator response, or CAR T-cell response rate, and g is the prey population density threshold at which predator behavior changes (e.g. fast-to-slow or slow-to-fast rates of killing). Functional responses model changes in predator hunting due to the prey density, generally defined with respect to some prey population threshold, here denoted as g. The population dependence on predator hunting behavior can also be interpreted as a handling time for distinguishing between time spent seeking prey, or recognizing cancer cells, and time spent consuming and attacking prey (25, 26, 19).

The three types of functional responses are graphed in Figure 3. In a Type I functional response, the predator response is constant for all prey population sizes. The interpretation of this response is that there are no differences in time or cost between all predator functions (searching and capture). In a Type II functional response the predator response is linear at low prey density (mirroring a Type I behavior) yet saturates at high prey density. Finally, in a Type III functional response the predator response is low at low prey densities, reflecting the potential for cancer cells to escape immune surveillance, yet again saturates at high prey densities, with a linear response at intermediate prey densities.

Importantly, the rational forms of Types II and III functional responses typically complicate determination of parameter values in conventional dynamical modeling. To reduce model complexity, we assume a significant level of effectiveness in CAR T-cell treatment such that the cancer cell population remains relatively low with respect to the functional response threshold, that is $x < g$, or $x/g < 1$. CAR T-cell effectiveness is demonstrated in Figure 1, where the control

cancer cell population is shown to achieve a maximum population of approximately 6.5 CI, while the treatment population of $E:T=1:4$ reaches a maximum population of approximately 2CI. The approximation condition permits the use of a binomial expansion about $x=0$ on the denominators for the Types II and III functional responses, resulting in,

$$\mathcal{R}(x)=\begin{cases} px & \text{Type I} \quad (12)\\ \dfrac{px}{g}\left(1-\dfrac{x}{g}+\dfrac{x^2}{g^2}+\displaystyle\sum_{j=3}^{\infty}(-1)^j\left(\dfrac{x}{g}\right)^j\right) & \text{Type II} \quad (13)\\ \dfrac{px^2}{g^2}\left(1-\dfrac{x^2}{g^2}+\dfrac{x^4}{g^4}+\displaystyle\sum_{j=3}^{\infty}(-1)^j\left(\dfrac{x^2}{g^2}\right)^j\right) & \text{Type III} \quad (14)\end{cases}$$

Further assuming that contributions to the functional response models of $\mathcal{O}(x^3/g^3)$ or greater are negligible, we terminate the expansions at $\mathcal{O}(x^2/g^2)$ to arrive at the following effective functional response models,

$$\mathcal{R}(x)=\begin{cases} px & \text{Type I} \quad (15)\\ \dfrac{px}{g}-\dfrac{px^2}{g^2} & \text{Type II} \quad (16)\\ \dfrac{px^2}{g^2} & \text{Type III} \quad (17)\end{cases}$$

It is important to highlight that the leading order term for the expansion for a Type II functional response is indistinguishable from a Type I functional response. This feature is reflected by the overlap in the graphs of the Type I and Type II responses presented in Figure 3, where the cancer cell population is small, $x\in[0,1]$ CI, compared to the value of $g=5$ CI. As the cancer cell population increases, the density dependence of the CAR T-cells starts to take effect as demonstrated by the parabolic contribution of the Type II response. In contrast to this, the expansions for functional responses of Types II and III are significantly unique from one another. Specifically, only expansions for Type II can lead to odd-powered terms in x, and although both expansions can express similar even-powered terms, they come with different concavities. That is, at small cancer populations the Type II functional response is characterized as a concave down parabola, while the Type III functional response is characterized as a concave up parabola. This difference regarding the positivity of the terms that are of second-order dependence in x corresponds to the different density dependent behaviors of the CAR T-cells at small cancer cell populations, specifically that Type II is a fast-to-slow response rate while Type III is a slow-to-fast response rate.

By performing the approximations used to derive Eqs.(16)-(17), and using truncated terms, we have reduced the complexity of the functional response terms. This step will simplify the process of model discovery. However, since this step assumes that the prey population remains small compared to the functional response threshold, the number of terms needed in Eqs.(13)-(14) for

accurate characterization of the system dynamics may vary as a result of experimental variation in the effector to target ratio of the CAR T-cells and the cancer cells. This variation in the effector to target ratio may also influence the structure of other interaction terms, specifically those pertaining to the single or paired binding dynamics.

2.2.3 CAR T-cell-cancer cell binding

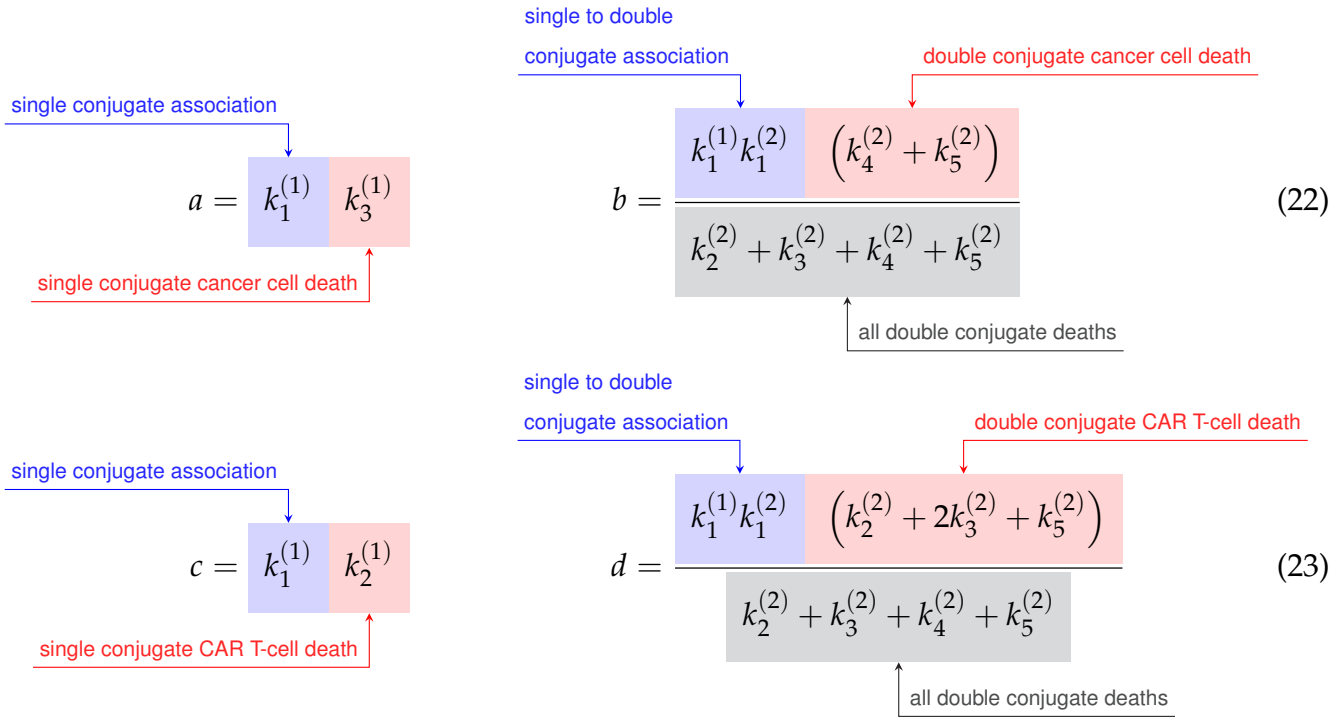
Cell binding models characterize the rates of formation and disassociation of conjugate pairs of species, also referred to as interaction molecules (Figure 4). These models historically are known as Hill-Langmuir functions for their originating studies in hemoglobin formation (52) and gas adsorption on material surfaces (53), yet perhaps are better known for their use in modeling enzyme reaction kinetics, or Michaelis-Menten kinetics (54). The same modeling principles have been extended to examine cell binding in T-cell and cancer cell interactions (2, 29, 30). An important challenge to the field of cancer immunotherapy modeling is characterizing higher-order cell binding dynamics. That is, the formation of conjugates that consist of multiple CAR T-cells attacking single cancer cells (Figure 4). These cancer cell-CAR T-cell conjugates are hypothesized to form as either a consequence of increased effector to target ratios or as a result of increased antigen density on target cells. As our experiment uses one single cell line with a high and uniform antigen expression level of IL13Rα2, we assume on average all cancer cells have approximately the same antigen density. We thus focus our attention to experimental variation in the effector to target ratios.

Following the work of Li et al. (30), we incorporate *fast irreversible* single and double cell binding into our generic model landscape. Here, fast binding implies that conjugate formation and dissociation occur quickly enough to maintain equilibrium in the conjugate populations, I_1 and I_2, such that $dI_1/dt = 0$ and $dI_2/dt = 0$. While irreversible means that all conjugate formation leads to death, or $k_{-1}^{(1)} = 0$ and $k_{-1}^{(2)} = 0$. These assumptions are consistent with the conditions of relatively higher effector to target ratios, or high antigen densities on target cells. They also imply that a mixture of conjugates and dissociates may exist, but that the dynamics happen such that the conjugate populations are fixed and do not change with time. Furthermore, we only consider the higher-order binding scenario of two CAR T-cells to one cancer cell. Solving for the contributions to the cancer and CAR T-cell populations due to binding dynamics results in,

$$\mathcal{B}_x(y) = \begin{cases} \dfrac{ay}{h} & \text{Single Binding} \quad (18) \\ \dfrac{ay + by^2}{h + ky} & \text{Double Binding} \quad (19) \end{cases}$$

$$\mathcal{B}_y(y) = \begin{cases} \dfrac{cy}{h} & \text{Single Binding} \quad (20) \\ \dfrac{cy + dy^2}{h + ky} & \text{Double Binding} \quad (21) \end{cases}$$

where the constants a, b, c, and d are defined in terms of the association rate constants, $k_1^{(1)}$ and $k_1^{(2)}$, and the death rate constants $k_2^{(1)}, k_3^{(1)}, k_2^{(2)}, k_3^{(2)}, k_4^{(2)}$, and $k_5^{(2)}$ from Figure 4 as follows,



$$a = k_1^{(1)} k_3^{(1)} \qquad b = \frac{k_1^{(1)} k_1^{(2)} \left(k_4^{(2)} + k_5^{(2)}\right)}{k_2^{(2)} + k_3^{(2)} + k_4^{(2)} + k_5^{(2)}} \tag{22}$$

$$c = k_1^{(1)} k_2^{(1)} \qquad d = \frac{k_1^{(1)} k_1^{(2)} \left(k_2^{(2)} + 2k_3^{(2)} + k_5^{(2)}\right)}{k_2^{(2)} + k_3^{(2)} + k_4^{(2)} + k_5^{(2)}} \tag{23}$$

Finally, the constant h is the sum of the single conjugate death rates, $h = k_2^{(1)} + k_3^{(1)}$, and the constant k is simply a renaming of the double conjugate association rate, $k_1^{(2)}$. As the variable renaming is admittedly complicated, the constants a, b, c, and d are defined to quickly identify end states of conjugate formation and have been located next to their corresponding interaction products in Figure 4.

The per-cancer cell binding models are graphed in Figure 5. Model parameter values used for the single and double cell binding models in Eqs. (18)-(27)) are: $a = 20$ CI^{-2}·hrs^{-2}, and $h = 16$ CI^{-1}·hrs^{-1} for single binding; and $a = 20$ CI^{-2}·hrs^{-2}, $b = 5$ CI^{-3}·hrs^{-2}, $h = 16$ CI^{-1}·hrs^{-1}, and $k = 2$ CI^{-2}·hrs^{-1} for double binding. We highlight that we are restricting ourselves to scenarios where increases in the CAR T-cell population during a given trial leads to increases in the likelihood of double binding, which results in super-linear increase of per-cancer cell binding. This restriction enforces concavity of the effective double cell binding model which we explore next. It is possible for the double binding model to exhibit a sub-linear increase in per-cancer cell antigen binding as the CAR T-cell population increases, and an overall decrease in cancer cell killing. However, this scenario does not agree with our experimental data of increased killing with increased effector-to-target ratios.

As with the functional response models, we are again faced with ODE model terms consisting of ratios of polynomials. However, potential differences between the rates of conjugate association and conjugate death can give rise to simplifications. If the product of the CAR T-cell population and the rate of forming double conjugates, ky, is small compared to the sum of the rates of single conjugate deaths, h, then $ky/h < 1$, and we can again perform a binomial expansion in the cell

binding denominators. A second way of interpreting this condition is to require the number of CAR T-cells to remain small compared to the ratio of the rate of double conjugate formation to the sum of the rates of single conjugate deaths, $y < h/k$. Performing the binomial expansion and truncating again at $\mathcal{O}(y^2)$ results in the following effective models of cell binding,

$$\mathcal{B}_x(y) = \begin{cases} \dfrac{ay}{h} & \text{Effective Single Binding} \quad (24) \\ \dfrac{ay}{h} + \dfrac{(bh-ak)y^2}{h^2} & \text{Effective Double Binding} \quad (25) \end{cases}$$

$$\mathcal{B}_y(y) = \begin{cases} \dfrac{cy}{h} & \text{Effective Single Binding} \quad (26) \\ \dfrac{cy}{h} + \dfrac{(dh-ck)y^2}{h^2} & \text{Effective Double Binding} \quad (27) \end{cases}$$

Here the effective double conjugate antigen binding model takes the form of the exact single conjugate binding model plus a correction due to double conjugate formation. Eqs. (25) and (27) are graphed in Figure 5, using the parameter values of $a = 20$ CI^{-2}·hrs^{-2}, $b = 2.75$ CI^{-3}·hrs^{-2}, $h = 16$ CI^{-1}·hrs^{-1}, and $k = 2$ CI^{-2}·hrs^{-1}. These values are chosen to demonstrate that the effective double binding model can accurately approximate both the exact single and double binding models for small CAR T-cell populations, $y < 1$ CI. Importantly, we note that if the parameter values b or d are sufficiently small, corresponding to low double conjugate CAR T-cell or cancer cell death rates, then the quadratic terms in Eqs. (25) and (27) will be negative, and the concavity of the effective double binding model deviate significantly from the exact model. This phenomenological consideration of the effective models sets an important constraint on the positivity of the coefficients for the quadratic terms in Eqs. (25) and (27), which we will revisit in Section 2.3.

2.2.4 Landscape of effective models

To gain a broader perspective of the overall form of our ODE models, we substitute the effective models for functional responses and antigen binding into Eqs. (1)-(2), arriving at,

$$\frac{dx}{dt} = \mathcal{G}_x(x) - \tilde{a}xy - \tilde{b}xy^2 \quad (28)$$

$$\frac{dy}{dt} = \mathcal{G}_y(y) \pm \alpha xy \pm \beta x^2 y - \tilde{c}xy^2 \quad (29)$$

where $\mathcal{G}$ again represents any of the potential growth-death models under consideration, $\tilde{a} = a/h$ and $\tilde{b} = (bh - ak)/h^2$ are redefined constants (both assumed to be positive) for the coefficients of the effective single and double binding models for the cancer cells, $\alpha xy = (p/g - c/h)xy$ and represents the combination of first order terms for CAR T-cell response and single binding, $\beta x^2 y = (p/g^2)xy$ and represents the potential second order term from the CAR T-cell response, and $\tilde{c}xy^2 = ((dh - ck)/h^2)xy^2$ represents the effective double binding model for the CAR T-cells.

We have explicitly used $\pm$ notation to indicate that we do not know *a priori* the signs for the xy and x^2y terms in Eq. (29), as these are determined by the relative contributions of Type I and first order Type II-like CAR T-cell responses and single antigen binding for the xy term, and whether or not second order Type II or first order Type III CAR T-cell response is occurring for the x^2y term. The benefit of the approach demonstrates the presence and/or sign conventions of the various model coefficients that we determine using the SINDy model discovery algorithm can be directly interpreted in terms of different underlying biological phenomena.

2.3 Model discovery

Our implementation of the model discovery techniques of dynamic mode decomposition and sparse identification of non-linear dynamics (SINDy) is performed in two stages. First is latent variable analysis, the extraction of the latent variable representing the CAR T-cell population from the time-varying cancer cell population. The second step is implementation of SINDy, whereupon the functional terms of the underlying models describing the dynamical system are determined.

2.3.1 Latent Variable Analysis

Despite having only measured the initial and final CAR T-cell populations, we can utilize latent variable analysis to infer the hidden CAR T-cell dynamics from the cancer cell dynamics. We do this using the delay coordinate embedding of Taken's theorem to reconstruct the attractor of the system that is known to exist in more dimensions than those measured (13, 15, 55). The first step in this approach is to assemble a Hankel matrix, $\mathbf{H}$, by stacking delayed time-series of the cancer cell measurements $x(t)$ as follows,

$$\mathbf{H} = \begin{bmatrix} x(t_1) & x(t_2) & x(t_3) & \cdots & x(t_{N-(m-1)\tau}) \\ x(t_{1+\tau}) & x(t_{2+\tau}) & x(t_{3+\tau}) & \cdots & x(t_{N-(m-2)\tau}) \\ x(t_{1+2\tau}) & x(t_{2+3\tau}) & x(t_{3+4\tau}) & \cdots & x(t_{N-(m-3)\tau}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x(t_{1+(m-1)\tau}) & x(t_{2+(m-2)\tau}) & x(t_{3+(m-3)\tau}) & \cdots & x(t_N) \end{bmatrix} \tag{30}$$

where τ, known as the embedding delay, represents the size of the time-delay we use, and m, known as the embedding dimension, represents both the number of rows that we assemble in the Hankel matrix and, importantly, the number of anticipated latent variables we expect to find.

To minimize the effects of experimental noise on the results of Taken's Theorem, we splined our cancer cell trajectories and re-sampled at the same experimental sampling rate of one measurement per 15 minutes. The function *smooth.spline* from the programming language **R** was used to perform the splining. This function uses cubic splines to approximate trajectories, with a penalty term to control for trajectory curvature. The number of knots used to spline each trajectory were determined by inspection, and are recorded in the analysis code available at `https://github.com/alexbbrummer/CART_SINDy`. Further details on the splining methods used are available in (56).

To determine optimal values for τ and m, we can use two separate formulae to inform the decisions (55). The optimal time delay is determined by the value of τ which minimizes the mutual information between measurements. This is done by dividing the interval $[x_{min}, x_{max}]$ into j equally sized partitions, and calculating the probability P_k that a measurement of the time

series is in the k^{th} partition, and the probability $P_{h,k}$ that a measurement x_i is in the h^{th} partition while the neighboring measurement $x_{i+\tau}$ is in the k^{th} partition. Mutual information is given by

$$I(\tau) = \sum_{h=1}^{j} \sum_{k=1}^{j} -P_{h,k}(\tau) \log \frac{P_{h,k}(\tau)}{P_h P_k}. \tag{31}$$

The optimal time-delay to use for a given time series is selected by finding the value of τ which results in the first minimum value in mutual information, or $\arg\min|_\tau\{I(\tau)\}$. A graph of mutual information versus time delay is presented in Figure 6**a**. For our cancer cell time series data, this optimal time delay value was found to be $\tau = 1$.

To determine the embedding dimension, m, we calculate the number of false nearest neighbors to a given measurement as the time series is embedded in successively greater dimensional spaces. This calculation is done to ensure that the attractor constructed from the latent variables remains smooth upon embedding. We perform the calculation iteratively by starting with a point $p(i)$ in an m-dimensional embedding, and identifying a neighboring point $p(j)$ such that the distance between $p(i)$ and $p(j)$ is less than a constant value typically chosen as the standard deviation of the data. Next, the normalized distance between the points $p(i)$ and $p(j)$ in the $m+1$-dimensional embedding is calculated using the following expression,

$$R_i = \frac{|x(t_{i+m\tau}) - x(t_{j+m\tau})|}{||p(i) - p(j)||} \tag{32}$$

R_i is calculated across the entire time series and iteratively for greater embeddings, $m = 1, 2, 3, \ldots$. False nearest neighbors are identified when $R_i > R_{threshold}$, where $R_{threshhold} = 10$ has been identified as satisfactory for most datasets (55). The ideal embedding dimension m is finally determined as that which results in a negligible fraction of false nearest neighbors. In Figure 6**b** we present the calculated fraction of false nearest neighbors versus embedding the dimension. For our dataset, we identified $m = 2$ as the ideal embedding dimension, indicating the existence of one latent variable that we interpret as representing the CAR T-cell population.

Using values of $\tau = 1$ for the time delay and $m = 2$ for the embedding dimension results in the following form of the Hankel matrix,

$$\mathbf{H} = \begin{bmatrix} x(t_1) & x(t_2) & x(t_3) & \cdots & x(t_{N-1}) \\ x(t_2) & x(t_3) & x(t_4) & \cdots & x(t_N) \end{bmatrix} \tag{33}$$

To extract the latent variable that represents the CAR T-cell time series, we perform a singular value decomposition of the Hankel matrix, $\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^*$ (15, 13). Here, the columns of $\mathbf{V}$ represent scaled and standardized versions of both the original data in the first column, and approximations of the latent data in the subsequent columns. As our experimental procedure measured the initial and final CAR T-cell populations, our final step was to re-scale and offset the latent CAR T-cell variable extracted from the second column of $\mathbf{V}$. In Figure 7 we present the measured cancer cells and CAR T-cells in addition to the discovered latent CAR T-cell time series for each effector to target ratio and duplicate trial.

2.3.2 Sparse identification of non-linear dynamics

SINDy is a data-driven methodology that discovers dynamical systems models through symbolic regression (15, 13). From a conceptual perspective, SINDy allows for the transformation of an analytical, first-order, non-linear dynamical systems model, expressed as

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t)) \tag{34}$$

to a linearized matrix-model, expressed as

$$\dot{\mathbf{X}} = \mathbf{\Theta}(\mathbf{X})\Xi \tag{35}$$

where $\dot{\mathbf{X}}$ are numerical time-derivatives of our measured data, $\Theta(\mathbf{X})$ is a library of candidate functions that may describe the data and is evaluated on the measured data, and Ξ consists of the coefficients for the model terms from $\Theta(\mathbf{X})$ that describe the time-varying data $\dot{\mathbf{X}}$. The objective of SINDy is to identify the sparsest version of Ξ, where sparsity is defined as the compromise between fewest number of non-zero terms with the greatest level of accuracy. In the context of our measurements for populations of cancer cells, $x(t)$, and CAR T-cells, $y(t)$, and the anticipated models for cell growth and interactions, $\dot{\mathbf{X}}$ takes the following form,

$$\dot{\mathbf{X}} = \begin{bmatrix} \dot{\mathbf{x}}^T(t_1) \\ \dot{\mathbf{x}}^T(t_2) \\ \vdots \\ \dot{\mathbf{x}}^T(t_N) \end{bmatrix} = \begin{bmatrix} \dot{x}(t_1) & \dot{y}(t_1) \\ \dot{x}(t_2) & \dot{y}(t_2) \\ \vdots & \vdots \\ \dot{x}(t_N) & \dot{y}(t_N) \end{bmatrix} \tag{36}$$

and $\mathbf{\Theta}(\mathbf{X})$ is expressed as,

$$\mathbf{\Theta}(\mathbf{X}) = \begin{bmatrix} | & | & | \\ \mathbf{X} & \mathbf{X}^2 & \mathbf{X}^3 \\ | & | & | \end{bmatrix} \tag{37}$$

$$\mathbf{\Theta}(\mathbf{X}) = \begin{bmatrix} x(t_1) & y(t_1) & x(t_1)^2 & x(t_1)y(t_1) & y(t_1)^2 & x(t_1)^3 & x(t_1)^2y(t_1) & x(t_1)y(t_1)^2 & y(t_1)^3 \\ x(t_2) & y(t_2) & x(t_2)^2 & x(t_2)y(t_2) & y(t_2)^2 & x(t_2)^3 & x(t_2)^2y(t_2) & x(t_2)y(t_2)^2 & y(t_2)^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x(t_N) & y(t_N) & x(t_N)^2 & x(t_N)y(t_N) & y(t_N)^2 & x(t_N)^3 & x(t_N)^2y(t_N) & x(t_N)y(t_N)^2 & y(t_N)^3 \end{bmatrix} \tag{38}$$

By solving the matrix-inverse problem in Eq. (35), we can find the column vectors Ξ that determine the coefficients for the model terms ξ that form the non-linear dynamical system best describing the measured data, in a least-squares sense. However, a simple least-squares implementation will result in a dense coefficient vector Ξ. We enforce sparsity of the coefficient vector Ξ through the method of sparse relaxed regularized regression (SR3) (57), where we seek optimization of the expression,

$$\min_{\Xi,W} \frac{1}{2}||\dot{\mathbf{X}} - \boldsymbol{\Theta}(\boldsymbol{\Xi})||^2 + \lambda R(\mathbf{W}) + \frac{1}{2\nu}||\boldsymbol{\Xi} - \mathbf{W}||^2 \tag{39}$$

where $\mathbf{W}$ is the relaxed coefficient matrix that approximates $\boldsymbol{\Xi}$, $R(\mathbf{W})$ is the regularization of $\mathbf{W}$, and ν and λ are hyper parameters that control how precisely $\mathbf{W}$ approximates $\boldsymbol{\Xi}$ and the strength of the regularization, respectively. For our problem, we chose to regularize under the $\ell 1$-norm with $\nu = 1 \times 10^{-5}$. To determine the value of λ, we followed the approach taken in (37) in which we repeat the analysis for a range of λ values from $\lambda \in [10^{-8}, 10^{1}]$ to calculate Pareto fronts between the root-mean-squared error between the measured and subsequently predicted values of $\mathbf{X}$ and the number of active terms from our library. In Figure 8 we present Pareto fronts for each of the experimental conditions for the varying effector to target ratios.

As discussed in Section 2.2, there are a variety of constraints we can expect for possible coefficients based on expected signs, or the absence of particular terms. An extension to SINDy allows for the incorporation of these constraints to ensure spurious terms are not discovered (58).

To make clear the constraints that were imposed, we can re-write Eq. (35) symbolically and in terms of the coefficients $\xi_{i,j}$ as,

$$\dot{x} = \xi_{1,1}x + \xi_{1,2}y + \xi_{1,3}x^2 + \xi_{1,4}xy + \xi_{1,5}y^2 + \xi_{1,6}x^3 + \xi_{1,7}x^2y + \xi_{1,8}xy^2 + \xi_{1,9}y^3 \tag{40}$$

$$\dot{y} = \xi_{2,1}x + \xi_{2,2}y + \xi_{2,3}x^2 + \xi_{2,4}xy + \xi_{2,5}y^2 + \xi_{2,6}x^3 + \xi_{2,7}x^2y + \xi_{2,8}xy^2 + \xi_{2,9}y^3 \tag{41}$$

Then, the constraints that are imposed as per the anticipated effective models from Section 2.2 are,

$$\xi_{1,2} = 0 \qquad \xi_{1,4} < 0 \qquad \xi_{1,5} = 0 \qquad \xi_{1,6} < 0 \tag{42}$$

$$\xi_{1,7} = 0 \qquad \xi_{1,8} < 0 \qquad \xi_{1,9} = 0 \qquad \xi_{2,1} = 0 \tag{43}$$

$$\xi_{2,3} = 0 \qquad \xi_{2,6} = 0 \qquad \xi_{2,8} < 0 \qquad \xi_{2,9} < 0 \tag{44}$$

while the other 6 coefficients in $\xi_{i,j}$ are left to freely vary.

Implementation of SINDy SR3 with constraints was performed using PySindy, a package designed for a wide array of implementations of the SINDy algorithm for spatio-temporal model discovery written in the programming language Python (16, 17). Included in the Supplementary Material are the associated datasets and Jupyter notebooks used for this study.

Finally, we highlight that the implementation of SINDy which we are relying on is designed specifically for explicit ordinary differential equations. An extension of SINDy exists for discovering ODEs with ratios of polynomials (37, 38), however this variation requires a significantly greater volume of data than that which we could collect. This is the underlying motivation behind our efforts to derive the effective models, thereby converting them into explicit ODEs and making effective usage of the volume of experimental data available by the study methods most usable for model discovery.

3 RESULTS

3.1 Discovered models and simulated comparison

Upon implementing SINDy on the CAR T-cell cancer cell killing data and performing the Pareto front analysis described in Section 2.3, we identified three distinct models describing the experimental data. Model selection is presented in Figure 8, where we present the tradeoffs between model complexity, represented by the number of activated library terms, and either threshold λ or the root-mean-squared-error between the measured data and simulated data for each identified model. Our examination of the Pareto fronts found models with eight terms for E:T of 1:4, and 1:8, and a six term model for an E:T of 1:20. Below we summarize each of these models and in relation to how well they predict the measured data in Figure 9. We synthesize the coefficients and associated model categories for growth in Table 1 and for the CAR T-cell functional response and cell binding in Table 2.

3.1.1 High E:T discovered model

For the E:T = 1:4 data, the SINDy-discovered model takes the following form,

$$\frac{dx}{dt} = 0.121x + 0.061x^2 - 0.018x^3 - 0.593xy^2 \tag{45}$$

$$\frac{dy}{dt} = 0.191y - 0.351y^2 + 0.035xy - 0.009x^2y \tag{46}$$

Factoring the terms related to single-species growth, we arrive at,

$$\frac{dx}{dt} = \underbrace{0.121x\left(1 - \frac{x}{4.792}\right)\left(\frac{x}{1.421} + 1\right)}_{\text{Weak Allee}} \underbrace{- 0.593xy^2}_{\text{Double Binding}} \tag{47}$$

$$\frac{dy}{dt} = \underbrace{0.191y\left(1 - \frac{y}{0.544}\right)}_{\text{Logistic}} \underbrace{+ 0.035xy - 0.009x^2y}_{\text{Type II Response}} \tag{48}$$

From Eqs. (47)-(48) we can interpret the discovered types of growth models and interactions. For cancer cell growth in Eq. (47), the observable structure indicates a weak Allee effect, with a growth rate of $\rho = 0.121$ hrs^{-1}, a carrying capacity of $K = 4.792$ CI, and an Allee constant of $A = 1.421$ CI. For the CAR T-cells we find a logistic growth model with growth rate $\rho = 0.191$ hrs^{-1} and carrying capacity $K = 0.544$ CI. From the coefficients of $\alpha = 0.051$ CI^{-1} hrs^{-1} on xy and $\beta = -0.009$ CI^{-2} hrs^{-1} on x^2y for the CAR T-cells, we can infer a Type II functional response as the signs are positive and negative, respectively. Finally, the presence of an xy^2 term in the cancer cells with a coefficient of $\tilde{b} = 0.063$ CI^{-2} hrs^{-1} indicates the occurrence of double binding, notably in the absence of both the xy term in the cancer cells and the xy^2 term in the CAR T-cells.

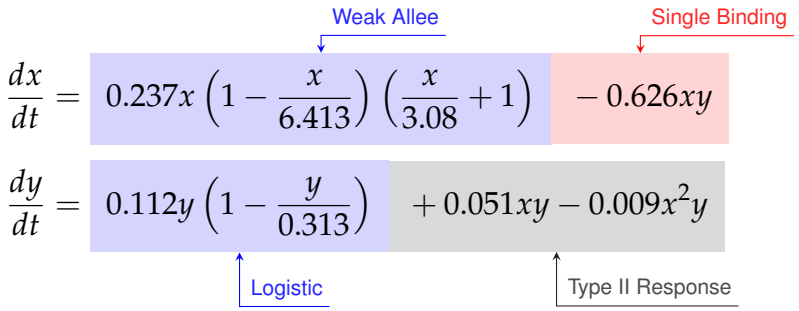
3.1.2 Medium E:T discovered model

The SINDy-discovered model for the E:T = 1:8 data takes the following form,

$$\frac{dx}{dt} = 0.237x + 0.04x^2 - 0.012x^3 - 0.626xy \tag{49}$$

$$\frac{dy}{dt} = 0.112y - 0.358y^2 + 0.051xy - 0.009x^2y \tag{50}$$

Factoring the terms related to single-species growth, we arrive at,



$$\frac{dx}{dt} = 0.237x\left(1 - \frac{x}{6.413}\right)\left(\frac{x}{3.08} + 1\right) - 0.626xy \tag{51}$$

$$\frac{dy}{dt} = 0.112y\left(1 - \frac{y}{0.313}\right) + 0.051xy - 0.009x^2y \tag{52}$$

The model discovered for medium E:T is largely similar to that at high E:T. A weak Allee effect in growth is observed for the cancer cells, with growth rate $\rho = 0.237$ hrs^{-1}, carrying capacity $K = 6.413$ CI, and Allee constant $A = 3.08$, while a logistic growth is observed for the CAR T-cells with growth rate $\rho = 0.112$ hrs^{-1} and carrying capacity $K = 0.313$ CI. We also observe a Type II CAR T-cell functional response, again indicated from the sign of the coefficients of $\alpha = 0.051$ CI^{-1} hrs^{-1} and $\beta = -0.01$ CI^{-2} hrs^{-1} on the xy and xy^2 terms being positive and negative, respectively. Unlike the high E:T scenario however, here we find evidence only of single binding from the sole presence of an xy term in the cancer cells with a coefficient of $\tilde{a} = -0.626$ CI^{-1} hrs^{-1}.

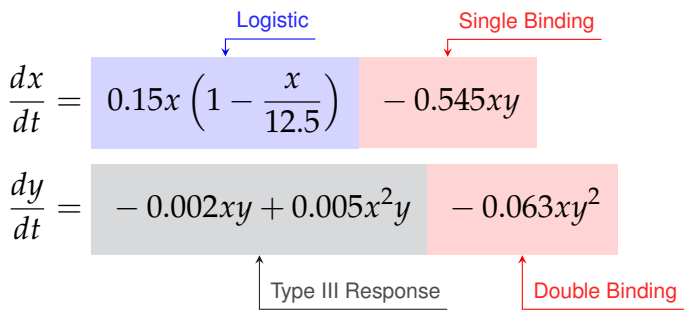
3.1.3 Low E:T discovered model

Finally, for the E:T = 1:20 data the discovered model is,

$$\frac{dx}{dt} = 0.150x - 0.012x^2 - 0.545xy \tag{53}$$

$$\frac{dy}{dt} = -0.002xy + 0.005x^2y - 0.063xy^2 \tag{54}$$

Factoring the terms related to single-species growth, we arrive at,



$$\frac{dx}{dt} = 0.15x\left(1 - \frac{x}{12.5}\right) - 0.545xy \tag{55}$$

$$\frac{dy}{dt} = -0.002xy + 0.005x^2y - 0.063xy^2 \tag{56}$$

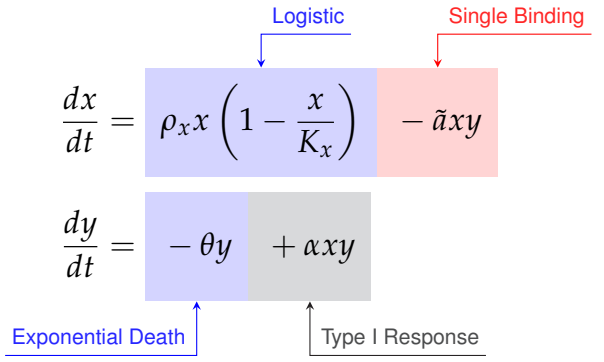
In this scenario we find significantly different growth and interaction models. The cancer cells

show logistic growth, with growth rate $\rho = 0.15$ hrs^{-1} and carrying capacity $K = 12.5$ CI, while the CAR T-cells have no growth model. This time, as the signs for the coefficients of $\alpha = -0.002$ CI^{-1} hrs^{-1} and $\beta = 0.005$ CI^{-2} hrs^{-1} on the xy and x^2y terms for the CAR T-cells are now negative and positive, respectively, we infer a Type III functional response. Interestingly, we find a mixture of indicators for both single binding and double binding. This comes from the presence of only the xy term in the cancer cell model with a coefficient of $\tilde{a} = -0.545$ CI^{-1} hrs^{-1}, and of an xy^2 term in the CAR T-cell model with a coefficient of $\tilde{c} = -0.063$ CI^{-2} hrs^{-1}.

All three E:T ratios of 1:4, 1:8, and 1:20 resulted in discovered models that accurately characterized the data, with root-mean-squared-errors of 0.02, 0.195, and 0.359, respectively. We highlight the discovery of consistent growth models of a weak Allee effect for the cancer cells and logistic growth for the CAR T-cells for the E:T ratios of 1:4 and 1:8. Importantly, the growth rates and carrying capacity for these scenarios were found to be comparable across E:T ratios. Interestingly, we observe a Type II functional response in the CAR T-cells functional response for both E:T = 1:4 and 1:8, and a transition to Type III for E:T = 1:20. Similarly, our discovered models indicate a transition from double to single binding as the E:T ratio changed from 1:4 to 1:8, and a model with mixed single and double binding terms was discovered for the E:T = 1:20.

3.2 Comparison with CARRGO model

We compared the *data first* model discovery methodology of SINDy against a traditional *model first* approach originally used to analyze and interpret the CAR T-cell killing dynamics (18, 24). The model previously studied is referred to as the CAR T-cell Response in GliOma (CARRGO) model, and is expressed as,



where we have expressed the parameter variables of the CARRGO model in terms of those used in the SINDy model for ease of comparison. From here we can see that the CARRGO model assumes logistic growth in the cancer cells, single binding between the cancer cells and CAR T-cells, a Type I functional response in the CAR T-cells, and exponential CAR T-cell death.

In its original and subsequent implementations, the CARRGO model demonstrated valuable utility in quantifying CAR T-cell killing dynamics when treating glioblastoma cell lines either alone (18), or in combination with the steroid dexamethasone (24). Inference of the underlying biological dynamics were made by examining how parameter values changed along gradients of E:T ratios or initial dexamethasone concentrations. This is in direct contrast to the SINDy methodology, where the discovery of different model terms provides insight into the underlying biological dynamics as a result of variation along the E:T gradient. Comparison of these two

models on the same data set provides further insight into the trade offs of *data first* versus *model first* approaches.

In Figure 10 are graphs of the best-fit versions of both the CARRGO model and SINDy discovered models for each E:T ratio. These fits were performed using the Levenberg-Marquadt optimization (LMO) algorithm, which requires initial guesses and bounds for each model parameter value. For the CARRGO model published parameter values were used for the starting guesses, while for the SINDy discovered models the discovered parameter values served as the guesses. Upper and lower bounds on the LMO search space were set at 80% and 120% of the originally identified parameter values, respectively, and are listed in the Supplementary Table. In Table 3 we present the model-fitting statistics for the reduced chi-squared, $\tilde{\chi}^2$, Akaike information criteria (AIC), and Bayesian information criteria (BIC) methods, as well as the parameters determined by LMO. Importantly, we note that fits were performed on data points representing averages and ranges for the two experimental trials at each E:T ratio from only the measured data.

We find that across the three statistical tests considered, the CARRGO model performs slightly better than the SINDy discovered models at E:T = 1:4 and E:T = 1:20, whereas the SINDy discovered model for E:T = 1:8 performed better than the CARRGO model (Table 3). Interestingly, the CARRGO model predictions for the CAR T-cell trajectories fail to intercept the final CAR T-cell values, where as the SINDy discovered models do. This result highlights a key difference between these two approaches, particularly that the SINDy approach required generating a time-series trajectory for the CAR T-cells that enforced interception with the final CAR T-cell measurement. Alternatively, traditional optimization methods like LMO weight each data point by the range of measurement uncertainty, allowing for the possibility of significant deviation from the final CAR T-cell measurements as long as such deviations can be compensated with better fitting elsewhere amongst the data.

Another essential difference between the CARRGO and SINDy predictions regarding the CAR T-cell trajectories is the CAR T-cell response at the high E:T ratio of E:T = 1:4. Specifically, the CARRGO model predicts that the CAR T-cells reach a maximum population exceeding the maximum population of cancer cells. This result has significant translational implications for CAR T-cell therapy related to patient immune response that we address in the discussion section.

Despite the noted differences, the overall similarities between the CARRGO and SINDy models is demonstrated by the order of magnitude agreement in most shared parameter values, specifically the cancer cell growth rate ρ_x, the cancer cell carrying capacity K_x, and the CAR T-cell functional response coefficient α for the specific scenarios of E:T = 1:4 and E:T = 1:8 (Table 3). Taken together, these results demonstrate significant value in the SINDy methodology when compared to established procedures for parameter estimation.

3.3 Model Stability

An important question in performing model discovery for dynamical systems is in relation to the overall stability. Automating the task of examining stability for every discovered model is challenging given the combination of symbolic computation with floating point coefficients. However, by predicting forward in time for each of the models and experimental replicates we can qualitatively characterize the stability (see Figure 9).

For the E:T = 1:4 scenario, both the data and model indicate complete cancer cell death, with the model accurately maintaining a cancer cell population of zero. We note that in several of the alternate discovered models produced by SINDy, the cancer cell population would become negative in the forward predicted regime. This unrealistic result can be used as an aide in ruling out alternative models.

For the E:T = 1:8 and 1:20 scenarios, both the data and models indicate cancer cell-CAR T-cell coexistence, with the forward predictions reaching non-oscillatory steady states. Despite the discovered models being the ones with the best accuracy, they all struggle to match the observed oscillatory frequency, particularly in the E:T = 1:20 scenario. These results demonstrate the capability of SINDy to discover models with variability in solution stability, a core feature of nonlinear dynamical systems.

3.4 Parameter Identifiability

To better understand the rarity of the discovered models and their respective coefficients, we examined histograms for the coefficients of each of the model terms along the Pareto fronts for each E:T ratio, presented in Figure 11. This approach allows us to qualitatively assess parameter identifiability by seeing the extent to which variability in coefficient values exists, and at the expense of prediction accuracy. For most active terms encountered, the coefficients corresponding to the selected models based on the Pareto front analysis were the most commonly occurring values until deactivation (elimination from discovered models). However, in a few situations we see that the coefficient values corresponding to the greatest model accuracy were relatively rare, and varied significantly as increasingly more terms were removed. This occurs in the coefficients for the x and xy^2 terms in the cancer cells for the E:T = 1:4 scenario in figure 11**a**, and the x and xy terms in the cancer cells for both the E:T = 1:8 and 1:20 scenarios in figure 11**b** and **c**. These terms were shown to be the final remaining active terms in discovered model, suggesting that they are capable of capturing the greatest extent of variation in our cancer cell-CAR T-cell killing data. Of note once again is that amongst these dominant interaction terms we see a transition from those indicative of double binding at high E:T ratios to single binding at medium and low E:T ratios.

4 DISCUSSION

We examined *in vitro* experimental CAR T-cell killing assay data for a human-derived glioblastoma cell line (Figure 1). From our results we infer transitions in the phenomenological killing behavior of the CAR T-cells as a consequence of varying their initial concentration compared to the cancer cells. Our discovered models predict that at high effector to target ratios (E:T = 1:4) the CAR T-cell levels respond according to a Type II functional response in which they survive and/or expand faster at low density, and slower at high density, and they predominantly form double binding conjugates with cancer cells prior to cell killing. At medium E:T ratios of E:T = 1:8 our discovered model again predicts the CAR T-cells undergoing a Type II functional response, but now forming only singly bound conjugates prior to cell killing. At low E:T ratios of E:T = 1:20 our discovered model predicts the CAR T-cells shift to a Type III functional response, in which they survive and/or expand slower at low density, and faster at high density. In this final scenario we find a mixture of single and double conjugate formation occurring. Finally, our discovered models predict the growth strategies of the cancer cells as being a weak Allee effect at high and medium E:T ratios, and logistic at low E:T ratios, while

the cancer cells are predicted to follow logistic growth for high and medium E:T ratios. Model coefficients used to deduce these results are found in Tables 1 and 2, and model simulations and forward predictions are shown in Figure 9.

A crucial result of this work is the comparison between the *data first* approach of SINDy to the traditional *model first* approach of CARRGO. Despite the discovered SINDy models having more degrees of freedom (i.e. mathematical terms) than the CARRGO model, both models were found to perform comparably as indicated in Figure 10 and Table 3. Yet, there are key differences regarding the interpretation of these two approaches. Traditional *model first* approaches like the CARRGO model assume a strict individual model that may exhibit variation in its coefficients or model parameters to reflect variation in the underlying biology or experimental conditions. On the other hand, one of the strengths of the *data first* approach of SINDy is that these coefficient variations can be shifted onto discovery of altogether different model terms. As we show, these different terms can have direct interpretations related to the underlying biology and dynamics. For example in (18), variation in the CAR T-cell response due to changes in the experimental E:T ratio could only be indicated through variation in the coefficients of the Type I functional response term, or the value of α in Eq. (58). Specifically, increases in α were interpreted as a high CAR T-cell response rate, or CAR T-cell expansion, and decreases in α were interpreted as a low response rate, or as CAR T-cell exhaustion. Whereas the SINDy model predicts entirely different CAR T-cell functional response terms, providing greater interpretation of these transitions in the CAR T dynamics and biology. Specifically, a Type II functional response at high and medium E:T, or a fast-to-slow CAR T-cell response rate, and a Type III functional response at low E:T, or a slow-to-fast CAR T-cell response that is again suggestive of exhaustion.

4.1 Interpreting Discovered Coefficients

We demonstrate the value of the effective model parameters for inferring underlying biology by considering the high E:T model presented in Eqs. (47)-(48). In this scenario, a Type II functional response in the CAR T-cells is deduced from the negative sign on β, corresponding to the concave down parabolic nature of the CAR T-cell functional response with fast proliferation at low cancer cell density and slow proliferation at high cancer cell density (Figure 3). The implication that cancer cell killing is induced by double binding of CAR T-cells to cancer cells comes from multiple terms. The most direct indicator is $\tilde{b} \neq 0$, where $\tilde{b} = (bh - ak)/k^2$ with bh/k representing the rate of cancer cell death from double conjugates, and a/k the rate of cancer cell death from single conjugates. Supporting indicators come from the positive sign on $\alpha = p/g - c/h$, suggesting that the CAR T-cell death rate from single conjugate formation, c/h is small compared to the leading order CAR T-cell response rate, p/g. Further evidence is in the inactivation of the xy term in the $\dot{x}(t)$ equation with coefficient $\tilde{a}$. Here, $\tilde{a} = a/h$ is the rate of cancer cell death from single conjugate formation, whose absence suggests that double binding formation is predominantly responsible for cancer cell death.

A similar analysis of model coefficients for the low and medium E:T ratio scenarios predicts a transition in the interactions between the CAR T-cells and cancer cells. Specifically, our approach predicts that the CAR T-cells form double conjugate pairs with high E:T ratios, then switch to single conjugate pairs at medium and low E:T ratios. Similarly, our results predict a transition in the functional response, indicating Type II functional responses in the CAR T-cells for high E:T ratios and Type III responses in the low E:T ratios. These transitions in detected model terms are

phenomenologically consistent with the interactions being dependent on CAR T-cell density, and highlight the hypothesis generating strength of *data first* model discovery techniques. Namely, the prediction of CAR T-cell killing dynamics being dependent on the relative abundance of CAR T-cells compared to cancer cells. We next present several opportunities for experimental testing of these model predictions.

4.2 Challenges and Limitations

A challengze to the implementation of SINDy is data sparsity. Despite having high temporal resolution of the cancer cell trajectories (1 measurement per 15 minutes), the CAR T-cell populations consisted of only the initial and final measurements. To resolve sparsity in the CAR T-cell levels, we used latent variable analysis to extract the CAR T-cell trajectory from an approximation to the attractor of the dynamical system as determined by the cancer cell trajectory. We note that in determining the dimensionality of the latent variable subspace, we selected an embedding dimension of $m = 2$ despite the appearance of further benefit in using an embedding dimension of $m = 3$, as indicated in Figure6**b**. This choice was made due to our experimental limitations in only having flow cytometry data for the CAR T-cells at the initial and final time points, and no further data with which to constrain any additional latent variables. The existence of a second latent variable, as suggested by the third embedding dimension, could be due to single or double binding conjugates if the reaction rates are sufficiently slow, or, alternatively, a biochemical secretion that is modulating the cancer cell and CAR T-cell interactions. Future experimental and modeling efforts may further illuminate the nature of this third state variable.

One potential limitation with latent variable analysis is that the trajectories retrieved through Taken's Theorem are not guaranteed to be unique, but rather will be diffeomorphic to the true latent variable. That is, subject to topological stretching or skewing, which translates to variation in discovered model coefficients. This effect can be seen in Bakarji et. al (59), where the coefficients of the latent variables discovered for the two-state, predator-prey model are not in precise agreement to those used in the original simulation. However, it is important to note that the model terms discovered by SINDy with this methodology are biologically insightful, even though the coefficients multiplying the discovered model terms on latent variables may be subject to variation. Importantly, we provide further experimental information for the latent CAR-T cell variable through bounding of the initial and final CAR-T cell trajectory with direct measurements. Likewise, we only discover terms which are structurally identifiable through model inversion, minimizing the potential for diffeomorphic skewing of CAR-T cell trajectories to be discovered from Taken's Theorem.

A second challenge is that our data in total consists of two trials for each effector-to-target ratio. While there exist SINDy implementations designed to discover models with ratios of polynomials, the approaches require prohibitively many experimental trials to ensure accuracy (37, 38). To resolve sparsity in the number of experimental trials, we derived effective interaction models of cancer cell and CAR T-cell dynamics from model ODE terms with ratios of the polynomials using binomial approximations. These effective interaction models allowed for the identification of multiple constraints on the library function space used in SINDy, and guided our inferential analysis of the discovered models.

4.3 Future Directions

To validate the hypothesized binding and functional response dynamics, we propose two potential experiments. Both experiments rely on similar initial conditions as those conducted for this study, but in one we propose the use of bright field microscopy to visually inspect CAR T-cell dynamics at different points in time and for the different E:T ratios. This approach ought to aide in identifying the relative abundance of binding types and functional responses. The second experiment would be to conduct endpoint analyses using flow cytometry to determine the population of CAR T-cells throughout the trajectory. This experiment would test the different CAR T-cell predictions from the CARRGO model and the SINDy models, most notably the predicted time to reach maximum CAR T-cell populations (Figure 10). Furthermore, targeted staining can provide information on the number of CAR T-cell generations and the ratio of helper T-cells (CD4+) to cytotoxic, killer T-cells (CD8+). These metrics may better inform the number of true effector cells responsible for killing cancer cells, allowing for more accurate characterization of CAR T-cell response.

Despite having high measurement sampling rates for the cancer cell population, the total experiment duration prevents the observation of highly periodic dynamics, a challenge to the standard implementation of SINDy in which observation windows generally span multiple periods of system dynamics. Thus we propose experimentally resolving the CAR T-cell trajectory to overcome this obstacle. This experiment additionally serves to test the validity of our latent variable analysis, which uses the cancer cell trajectory to predict the CAR T-cell trajectory as presented in Figure 7. Future experiments will also extend this analysis to include other CAR designs, including evaluating the impact of costimulatory signaling, CAR affinity and target density on modeling of CAR T-cell killing dynamics.

4.4 Clinical Applications

The clinical relevance of the *data first* framework is in the domain of precision medicine. The approach naturally caters to *in situ* monitoring of patient response to therapy and forecasting future trajectories. An open question in this field is determining the sufficient number of early measurements necessary for accurate forecasting, and quantifying the extent of reliable forward prediction. This type of application falls under the field of control theory, in which real-time measurements for systems such as navigation, fluid dynamics and disease monitoring can inform model-based interventions (15). Control theory has been identified as a key tool in achieving optimized individual treatment outcomes, yet challenges are ever-present in parsimonious model selection. The SINDy methodology may help streamline and simplify the model selection process, while simultaneously incorporating control theory methods for treatment optimization. As an example related to the experiments considered here, one could envision a therapeutic intervention to administer more CAR T-cells in the low E:T ratio of 1:20 as soon as the Type III functional response and single binding dynamics are predicted in a patient. This intervention would serve to push the dynamics of the patients immune response into the double biding and Type II response regime, thereby improving therapeutic efficacy.

5 CONCLUSIONS

In this work we present the first, to our knowledge, application of the sparse identification of non-linear dynamics (SINDy) methodology to a real biological system. We used SINDy with highly time-resolved experimental data to discover biological mechanisms underlying CAR T-cell-cancer cell killing dynamics. Our implementation highlights the hypothesis generating potential of data-driven model discovery and illuminates challenges for future extensions and applications. To overcome challenges related to data limitation, we utilized latent variable analysis to construct the trajectory of the CAR T-cells, and we implemented binomial expansions to simplify specific model terms. Our results predict key mechanisms and transitions in the interaction dynamics between the CAR T-cells and cancer cells under different experimental conditions that may be encountered in the application of these therapies in human patients. Specifically, we identified transitions from double CAR T-cell binding to single CAR T-cell binding, and from fast-to-slow CAR T-cell responses (Type II) to slow-to-fast responses (Type III). Both transitions occur as a result of decreasing the relative abundances of CAR T-cells to cancer cells (initial E:T ratios). Importantly, these results demonstrate the potential for *data first* model discovery methods to provide deeper insight into the underlying dynamics and biology than *model first* approaches, and offer a new avenue for integrating predictive modeling into precision medicine and cancer therapy by an improved mechanistic understanding of cancer progression and efficacy of CAR T-cell therapy.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

AUTHOR CONTRIBUTIONS

Conceptualization, AB and RR; methodology, AB, RW, VA, HC and RR; validation, AB, RR; formal analysis, AB; investigation, AB; resources, MG, CB, and RR; data curation, AB; writing—original draft preparation, AB; writing—review and editing, AB, AX, AW, VA, HC, MG, CB, RR; visualization, AB, AX, RW, VA, RR; supervision, CB and RR; project administration, AB and RR; funding acquisition, CB and RR All authors have read and agreed to the published version of the manuscript.

FUNDING

Research reported in this work was supported by the National Cancer Institute of the National Institutes of Health under grant numbers R01CA254271 (CB), R01NS115971 (CB, RR), and P30CA033572, the Marcus Foundation, and the California Institute of Regenerative Medicine (CIRM) under CLIN2-10248 (CB).

SUPPLEMENTAL MATERIALS

Supplementary Table 1 - Parameter seed values for Levenberg-Marquardt Optimization of CARRGO and SINDy growth-death model terms.

Supplementary Table 2 - Parameter seed values for Levenberg-Marquardt Optimization of CARRGO and SINDy interaction model terms.

DATA AVAILABILITY STATEMENT

The data presented in this study and Python code used to analyze data and generate figures are openly available from the following online location `https://github.com/alexbbrummer/CART_SINDy`.

REFERENCES

1 .Rockne RC, Hawkins-Daarud A, Swanson KR, Sluka JP, Glazier JA, Macklin P, et al. The 2019 mathematical oncology roadmap. *Physical biology* **16** (2019) 041005.

2 .Kuznetsov VA, Makalkin IA, Taylor MA, Perelson AS. Nonlinear dynamics of immunogenic tumors: parameter estimation and global bifurcation analysis. *Bulletin of mathematical biology* **56** (1994) 295–321.

3 .de Pillis LG, Radunskaya AE, Wiseman CL. A validated mathematical model of cell-mediated immune response to tumor growth. *Cancer research* **65** (2005) 7950–7958.

4 .Roy M, Finley SD. Computational model predicts the effects of targeting cellular metabolism in pancreatic cancer. *Frontiers in Physiology* **8** (2017) 217. doi:10.3389/fphys.2017.00217.

5 .Stein AM, Grupp SA, Levine JE, Laetsch TW, Pulsipher MA, Boyer MW, et al. Tisagenlecleucel model-based cellular kinetic analysis of chimeric antigen receptor–t cells. *CPT: Pharmacometrics & Systems Pharmacology* **8** (2019) 285–295.

6 .West J, You L, Zhang J, Gatenby RA, Brown JS, Newton PK, et al. Towards multidrug adaptive therapy. *Cancer Research* **80** (2020) 1578–1589. doi:10.1158/0008-5472.CAN-19-2669.

7 .Kimmel GJ, Locke FL, Altrock PM. The roles of t cell competition and stochastic extinction events in chimeric antigen receptor t cell therapy. *Proceedings of the Royal Society B* **288** (2021) 20210229.

8 .Owens K, Bozic I. Modeling car t-cell therapy with patient preconditioning. *Bulletin of Mathematical Biology* **83** (2021) 1–36.

9 .Mascheroni P, Savvopoulos S, Alfonso JCL, Meyer-Hermann M, Hatzikirou H. Improving personalized tumor growth predictions using a bayesian combination of mechanistic modeling and machine learning. *Communications medicine* **1** (2021) 1–14.

10 .Lambin P, Leijenaar RT, Deist TM, Peerlings J, De Jong EE, Van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **14** (2017) 749.

11 .Zhou M, Leung A, Echegaray S, Gentles A, Shrager JB, Jensen KC, et al. Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. *Radiology* **286** (2018) 307–315.

12 .Kamb M, Kaiser E, Brunton SL, Kutz JN. Time-delay observables for koopman: Theory and applications. *SIAM Journal on Applied Dynamical Systems* **19** (2020) 886–917.

13 .Brunton SL, Budišić M, Kaiser E, Kutz JN. Modern koopman theory for dynamical systems. *SIAM Review* **64** (2022) 229–340. doi:10.1137/21M1401243.

14 .Frankhouser DE, O'Meally D, Branciamore S, Uechi L, Zhang L, Chen YC, et al. Dynamic patterns of microrna expression during acute myeloid leukemia state-transition. *Science advances* **8** (2022) eabj1664.

15 .Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences* **113** (2016) 3932–3937.

16 .de Silva BM, Champion K, Quade M, Loiseau JC, Kutz JN, Brunton SL. Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data. *Journal of Open Source Software* **5** (2020) 2104. doi:10.21105/joss.02104.

17 .Kaptanoglu AA, de Silva BM, Fasel U, Kaheman K, Goldschmidt AJ, Callaham J, et al. Pysindy: A comprehensive python package for robust sparse system identification. *Journal of Open Source Software* **7** (2022) 3994. doi:10.21105/joss.03994.

18 .Sahoo P, Yang X, Abler D, Maestrini D, Adhikarla V, Frankhouser D, et al. Mathematical deconvolution of car t-cell proliferation and exhaustion from real-time killing assay data. *Journal of the Royal Society Interface* **17** (2020) 20190734.

19 .Hamilton PT, Anholt BR, Nelson BH. Tumour immunotherapy: lessons from predator–prey theory. *Nature Reviews Immunology* (2022) 1–11.

20 .Lotka AJ. *Elements of physical biology* (Williams & Wilkins) (1925).

21 .Volterra V. Variazioni e fluttazioni del numero d'individui in specie animali conviventi. *Mem. Acad. Lincei Roma* (1926) 31–113.

22 .Kareva I, Luddy KA, O'Farrelly C, Gatenby RA, Brown JS. Predator-prey in tumor-immune interactions: A wrong model or just an incomplete one? *Frontiers in Immunology* (2021) 3391.

23 .Adhikarla V, Awuah D, Brummer AB, Caserta E, Krishnan A, Pichiorri F, et al. A mathematical modeling approach for targeted radionuclide and chimeric antigen receptor t cell combination therapy. *Cancers* **13** (2021). doi:10.3390/cancers13205171.

24 .Brummer AB, Yang X, Ma E, Gutova M, Brown CE, Rockne RC. Dose-dependent thresholds of dexamethasone destabilize car t-cell treatment efficacy. *PLOS Computational Biology* **18** (2022) e1009504.

25 .Holling CS. The components of predation as revealed by a study of small-mammal predation of the european pine sawfly1. *The Canadian Entomologist* **91** (1959) 293–320.

26 .Gatenby RA. Population ecology issues in tumor growth. *Cancer Research* **51** (1991) 2542–2547.

27 .Böttger K, Hatzikirou H, Voss-Böhme A, Cavalcanti-Adam EA, Herrero MA, Deutsch A. An emerging allee effect is critical for tumor initiation and persistence. *PLOS Computational Biology* **11** (2015) 1–14. doi:10.1371/journal.pcbi.1004366.

28 .Johnson KE, Howard G, Mo W, Strasser MK, Lima EABF, Huang S, et al. Cancer cell population growth kinetics at low densities deviate from the exponential growth model and suggest an allee effect. *PLOS Biology* **17** (2019) 1–29. doi:10.1371/journal.pbio.3000399.

29 .Chaudhury A, Zhu X, Chu L, Goliaei A, June CH, Kearns JD, et al. Chimeric antigen receptor t cell therapies: A review of cellular kinetic-pharmacodynamic modeling approaches. *The Journal of Clinical Pharmacology* **60** (2020) S147–S159.

30 .Li R, Sahoo P, Wang D, Brown CE, Rockne RC, Cho H. Modeling interaction of glioma cells and car t-cells considering multiple car t-cells bindings. *arXiv* (2022).

31 .Gábor A, Banga JR. Robust and efficient parameter estimation in dynamic models of biological systems. *BMC systems biology* **9** (2015) 1–25.

32 .Mitra ED, Hlavacek WS. Parameter estimation and uncertainty quantification for systems biology models. *Current Opinion in Systems Biology* **18** (2019) 9–18. doi:10.1016/j.coisb.2019.10.006.

33 .Schmid PJ. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics* **656** (2010) 5–28.

34 .Yadav N, Ravela S, Ganguly AR. Machine learning for robust identification of complex nonlinear dynamical systems: Applications to earth systems modeling (2020). doi:10.48550/ARXIV.2008.05590.

35 .Alves EP, Fiuza F. Data-driven discovery of reduced plasma physics models from fully kinetic simulations. *Phys. Rev. Research* **4** (2022) 033192. doi:10.1103/PhysRevResearch.4.033192.

36 .Ren Y, Adams C, Melz T. Uncertainty analysis and experimental validation of identifying the governing equation of an oscillator using sparse regression. *Applied Sciences* **12** (2022). doi:10.3390/app12020747.

37 .Mangan NM, Brunton SL, Proctor JL, Kutz JN. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* **2** (2016) 52–63.

38 .Kaheman K, Kutz JN, Brunton SL. Sindy-pi: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proceedings of the Royal Society A* **476** (2020) 20200279.

39 .Kaiser E, Kutz JN, Brunton SL. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A* **474** (2018) 20180335.

40 .Messenger DA, Bortz DM. Weak sindy: Galerkin-based data-driven model selection. *Multiscale Modeling & Simulation* **19** (2021) 1474–1497. doi:10.1137/20M1343166.

41 .Chen Z, Liu Y, Sun H. Physics-informed learning of governing equations from scarce data. *Nature communications* **12** (2021) 1–13.

42 .Brown CE, Starr R, Aguilar B, Shami AF, Martinez C, D'Apuzzo M, et al. Stem-like tumor-initiating cells isolated from il13rα2 expressing gliomas are targeted and killed by il13-zetakine–redirected t cells. *Clinical Cancer Research* **18** (2012) 2199–2209.

43 .Brown CE, Alizadeh D, Starr R, Weng L, Wagner JR, Naranjo A, et al. Regression of glioblastoma after chimeric antigen receptor t-cell therapy. *New England Journal of Medicine* **375** (2016) 2561–2569.

44 .Brown CE, Aguilar B, Starr R, Yang X, Chang WC, Weng L, et al. Optimization of il13rα2-targeted chimeric antigen receptor t cells for improved anti-tumor efficacy against glioblastoma. *Molecular Therapy* **26** (2018) 31–44.

45 .Moniri MR, Young A, Reinheimer K, Rayat J, Dai LJ, Warnock GL. Dynamic assessment of cell viability, proliferation and migration using real time cell analyzer system (rtca). *Cytotechnology* **67** (2015) 379–386.

46 .Xing JZ, Zhu L, Gabos S, Xie L. Microelectronic cell sensor assay for detection of cytotoxicity and prediction of acute toxicity. *Toxicology in vitro* **20** (2006) 995–1004.

47 .Chiu CH, Lei KF, Yeh WL, Chen P, Chan YS, Hsu KY, et al. Comparison between xcelligence biosensor technology and conventional cell culture system for real-time monitoring human tenocytes proliferation and drugs cytotoxicity screening. *Journal of orthopaedic surgery and research* **12** (2017) 1–13.

48 .Hanke M. *Conjugate gradient type methods for ill-posed problems* (Longman Scientific and Technical) (1995).

49 .Fadai NT, Simpson MJ. Population dynamics with threshold effects give rise to a diverse family of allee effects. *Bulletin of mathematical biology* **82** (2020) 1–22.

50 .Fadai NT, Johnston ST, Simpson MJ. Unpacking the allee effect: determining individual-level mechanisms that drive global population dynamics. *Proceedings of the Royal Society A* **476** (2020) 20200350.

51 .Cosner C, Rodriguez N. The effect of directed movement on the strong allee effect. *SIAM Journal on Applied Mathematics* **81** (2021) 407–433. doi:10.1137/20M1330178.

52 .Hill AV. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *Proceedings of the Physiological Society* **40** (1910) i–vii. doi:https://doi.org/10.1113/jphysiol.1910.sp001386.

53 .Langmuir I. The adsorption of gases on plane surfaces on glass, mica and platinum. *Journal of the American Chemical Society* **40** (1918) 1361–1403. doi:10.1021/ja02242a004.

54 .Michaelis L, Menten ML, et al. Die kinetik der invertinwirkung. *Biochemische Zeitschrift* **49** (1913) 352.

55 .Kodba S, Perc M, Marhl M. Detecting chaos from a time series. *European journal of physics* **26** (2004) 205.

56 .Brummer AB, Hunt D, Savage V. Improving blood vessel tortuosity measurements via highly sampled numerical integration of the frenet-serret equations. *IEEE Transactions on Medical Imaging* **40** (2021) 297–309. doi:10.1109/TMI.2020.3025467.

57 .Zheng P, Askham T, Brunton SL, Kutz JN, Aravkin AY. A unified framework for sparse relaxed regularized regression: Sr3. *IEEE Access* **7** (2018) 1404–1423.

58 .Champion K, Zheng P, Aravkin AY, Brunton SL, Kutz JN. A unified sparse optimization framework to learn parsimonious physics-informed models from data. *IEEE Access* **8** (2020) 169259–169271. doi:10.1109/ACCESS.2020.3023625.

59 .Bakarji J, Champion K, Kutz JN, Brunton SL. Discovering governing equations from partial measurements with deep delay autoencoders (2022). doi:10.48550/ARXIV.2201.05136.

FIGURE CAPTIONS

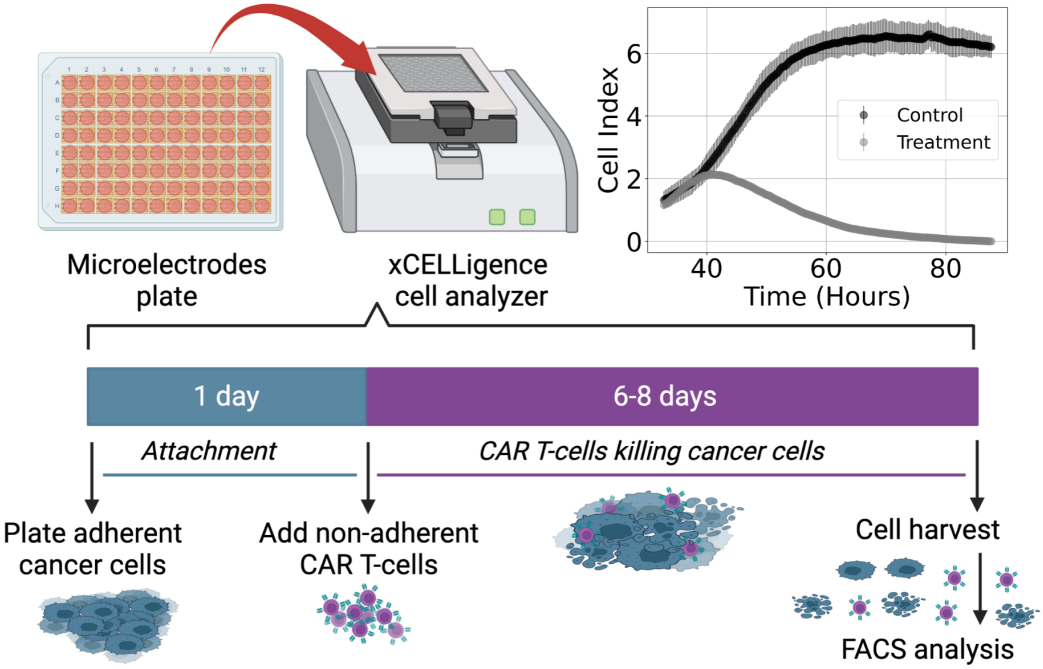


Figure 1. Diagram of experimental procedure highlighting use of microelectrode plates in an xCELLigence cell analyzer system and sample Cell Index (CI) measurements for control and treatment groups (E:T = 1:4). This system utilizes real-time voltage measurements to determine CI values representative of the adherent cancer cell population as a function of time. CAR T-cells are added following 24 hours of cancer cell expansion and attachment. After 6-8 days of monitoring the cancer cell growth and death dynamics, cells are harvested and enumerated using flow cytometry.

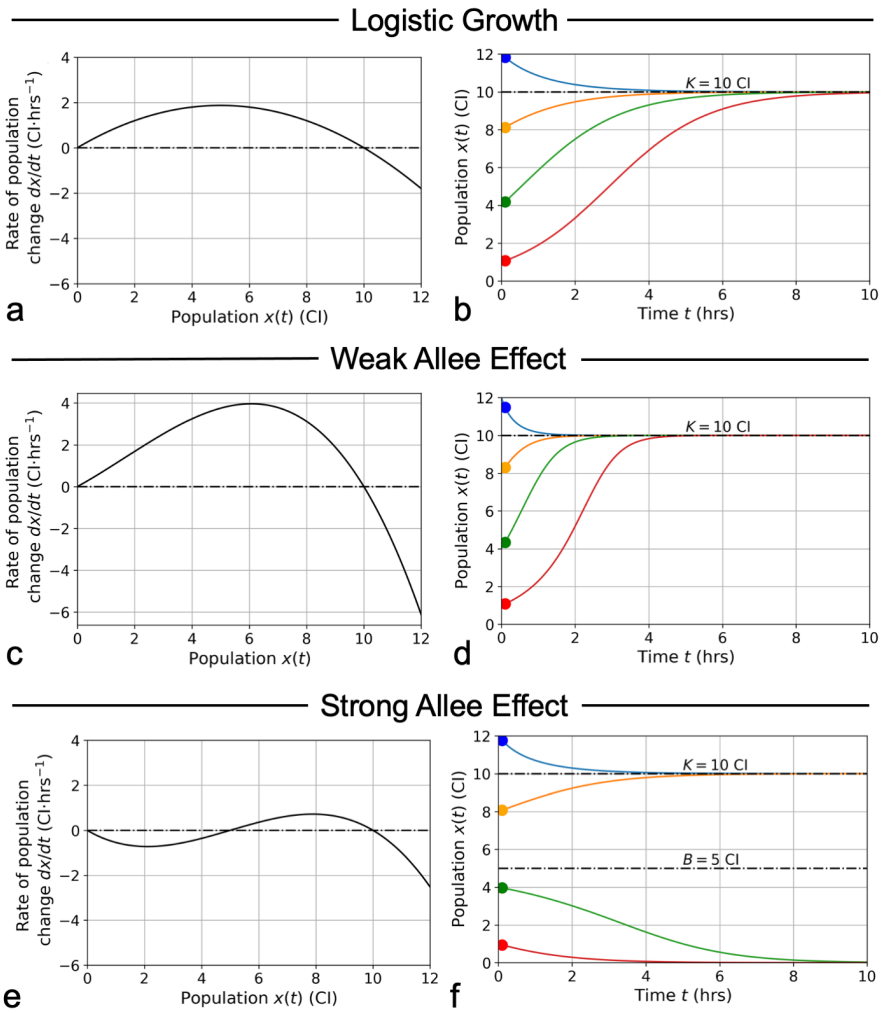


Figure 2. Conceptual graphs of rates of change in cancer cell populations versus population size (left panels) and population size versus time (right panels) for the three growth models presented in Eqs. (3)-(5): logistic growth **a** and **b**; weak Allee effect **c** and **d**; strong Allee effect **e** and **f**. Model parameter values are: $\rho = 0.75$ hrs^{-1}, $K = 10$ CI, $A = 5$ CI, and $B = 5$ CI. Colors correspond to different initial cancer cell seeding conditions which are the same for each model in a cancer cell only scenario (blue = 12CI, orange = 8CI, green = 4CI, red = 1CI).

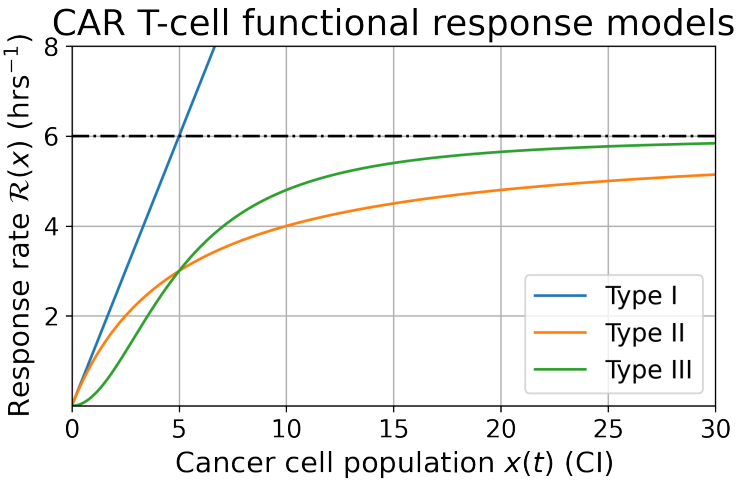


Figure 3. Functional responses (per-CAR T-cell response rates) versus cancer cell population. Model parameters for functional responses are: $p = 6/5$ CI^{-1}·hrs^{-1} for Type I; $p = 6$ CI^{-1}·hrs^{-1} and $g = 5$ CI for Types II and III. Note overlap of Types I and II functional responses for $x < 1$ CI, and distinct differences in concavity between Types II (negative) and III (positive) for $x < 2$ CI. These characteristics correspond to Type I and Type II functional responses being indistinguishable at low cancer cell populations, and Type II and Type III being differentiated by fast-then-slow response rates (Type II) versus slow-then-fast response rates (Type III).

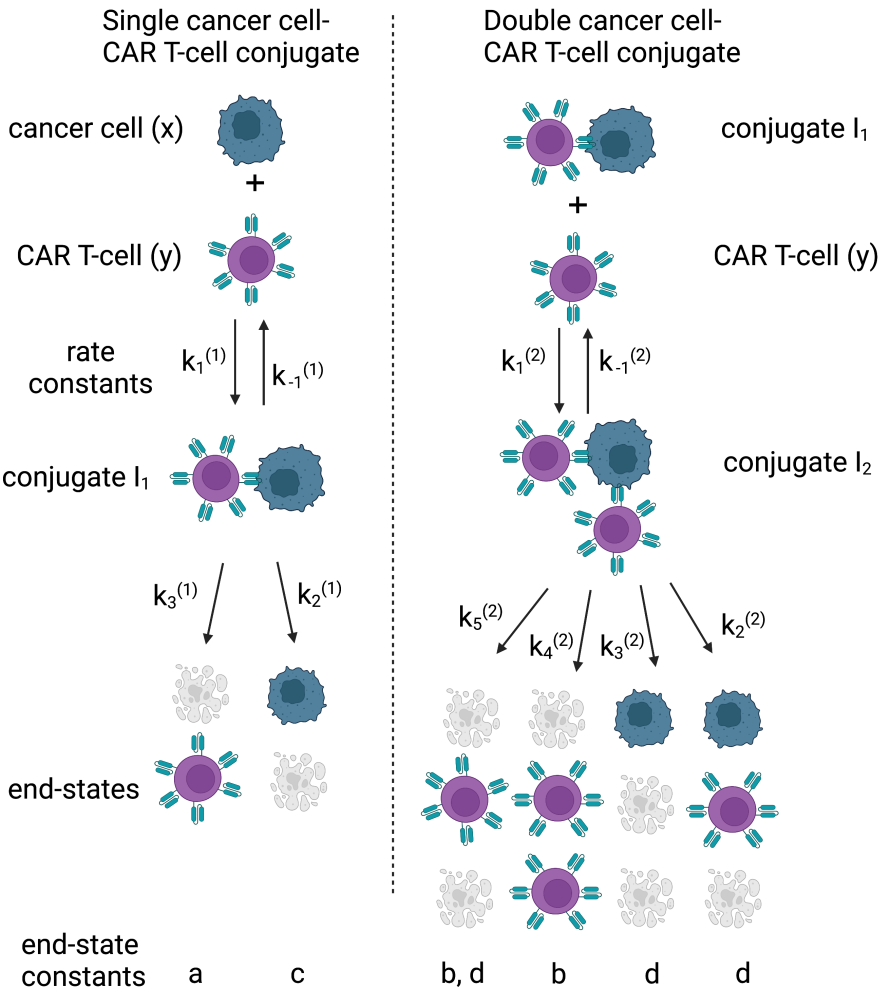


Figure 4. Compartmental model for single and double cell binding kinetics. Expressions for how rate constants combine to contribute to the growth or death of the cancer cell and CAR T-cell populations are presented in Eqs. (18)-(23). See (30) for further development and analysis of the cell binding model.

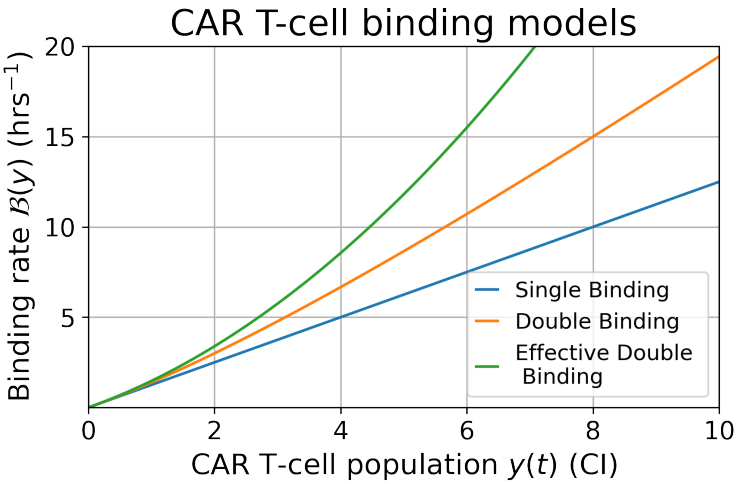


Figure 5. Binding rates (per-cancer cell CAR T-cell binding) versus CAR T-cell population for the single binding, double binding, and effective double binding models in Eqs. (18)-(21), (25), and (27). Model parameters for antigen bindings are: $a = 20$ CI^{-2}·hrs^{-2} and $h = 16$ CI^{-1}·hrs^{-1} for single binding; $a = 20$ CI^{-2}·hrs^{-2}, $b = 5$ CI^{-3}·hrs^{-2}, $h = 16$ CI^{-1}·hrs^{-1}, and $k = 2$ CI^{-2}·hrs^{-1} for double binding; and $a = 20$ CI^{-2}·hrs^{-2}, $b = 2.75$ CI^{-3}·hrs^{-2}, $h = 16$ CI^{-1}·hrs^{-1}, and $k = 2$ CI^{-2}·hrs^{-1} for effective double binding. These parameter values were chosen to highlight how well the effective double binding model can approximate both the single and double binding models at low CAR T-cell population values, $y < 1$ CI. Note that since the original double binding model in this scenario is concave-up, the effective double binding model parameters should be chosen to match concavity. This requirement sets a positivity constraint on the quadratic term in Eqs. (25) and (27).

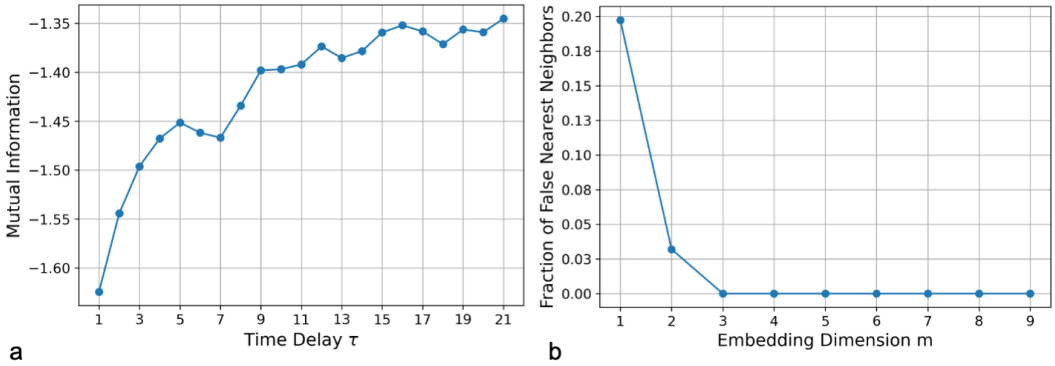


Figure 6. Graphs for determining **a** the ideal time delay, τ, by examining mutual information and **b** the ideal embedding dimension, m, by examining the fraction of false nearest neighbors. These methods are explained in detail in (55). For our data, the optimal time delay found was $\tau = 1$ and the optimal embedding dimension $m = 2$.

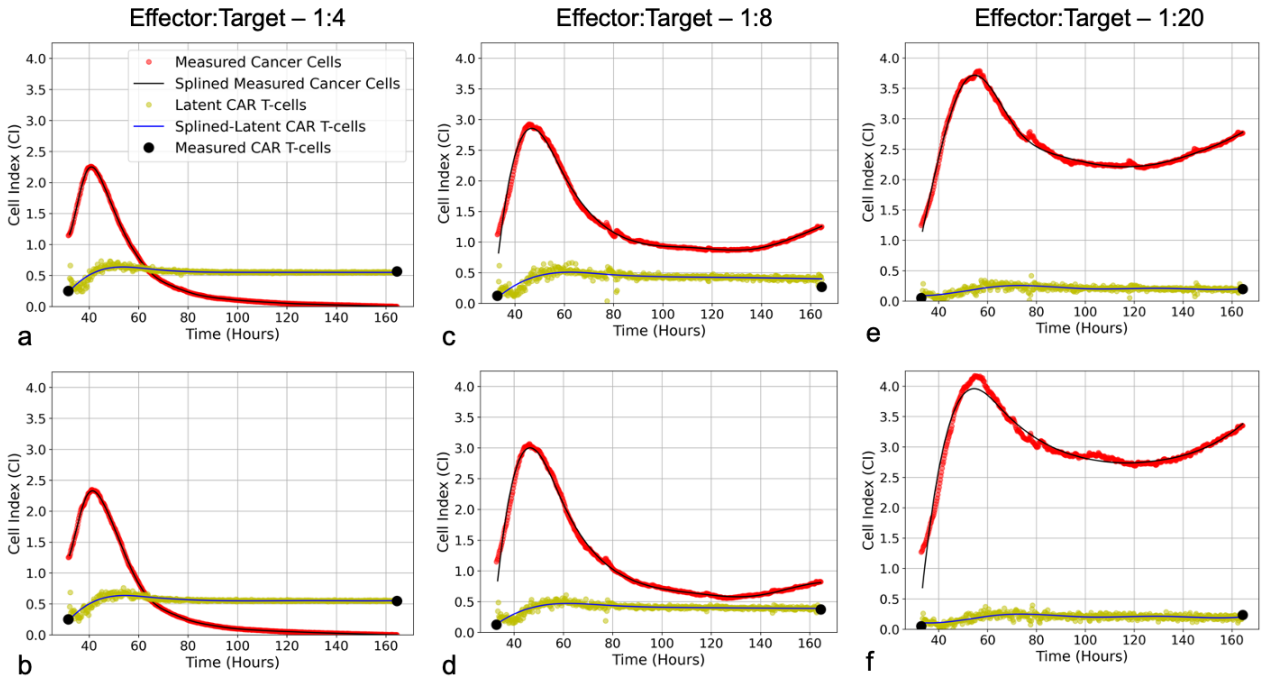


Figure 7. Latent variable analysis results. **a-b** are replicates 1 and 2 for the high E:T ratio, **c-d** are replicates 1 and 2 for the medium E:T ratio, and **e-f** are replicates 1 and 2 for the low E:T ratio. Presented for each trial are the cancer cell index measurements from the xCELLigence machine in red, overlaid with the splined measurements for the cancer cells in black; the two endpoint measurements for the CAR T-cell levels enumerated by flow cytometry in black, with the CAR T-cell population trajectory as determined by latent variable analysis in yellow, overlaid with the splined CAR T-cell trajectory in blue. Note that despite the CAR T-cell populations being measured with flow cytometry, we have converted levels to units of Cell Index for ease of comparison with the cancer cells, using a conversion factor of 1CI $\approx 10,000$ cells.

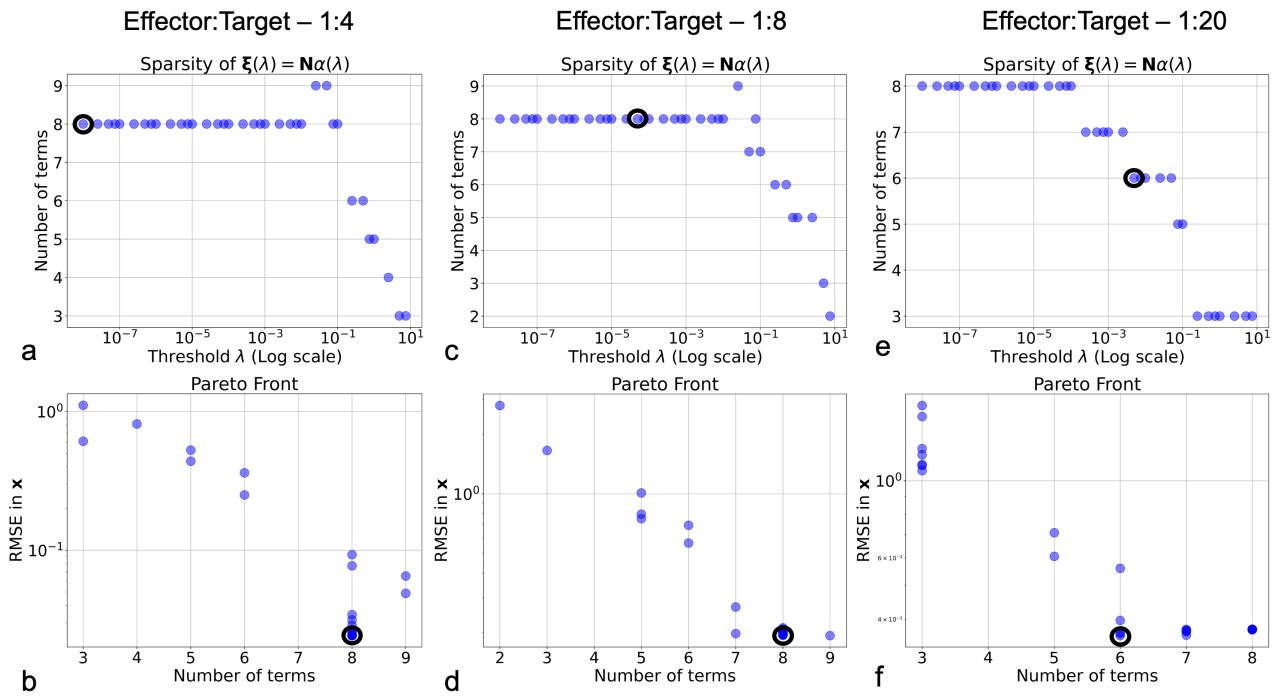


Figure 8. Pareto front analysis of system sparsity. **a-b**: high E:T, **c-d** medium E:T. **a** and **c** sparsity (number of discovered terms) versus threshold λ and **b** and **c** root-mean-squared-error (RMSE) versus number of terms. Note that the RMSE values were calculated using the discovered models and the splined measurements. Selected models are represented by purple circles.

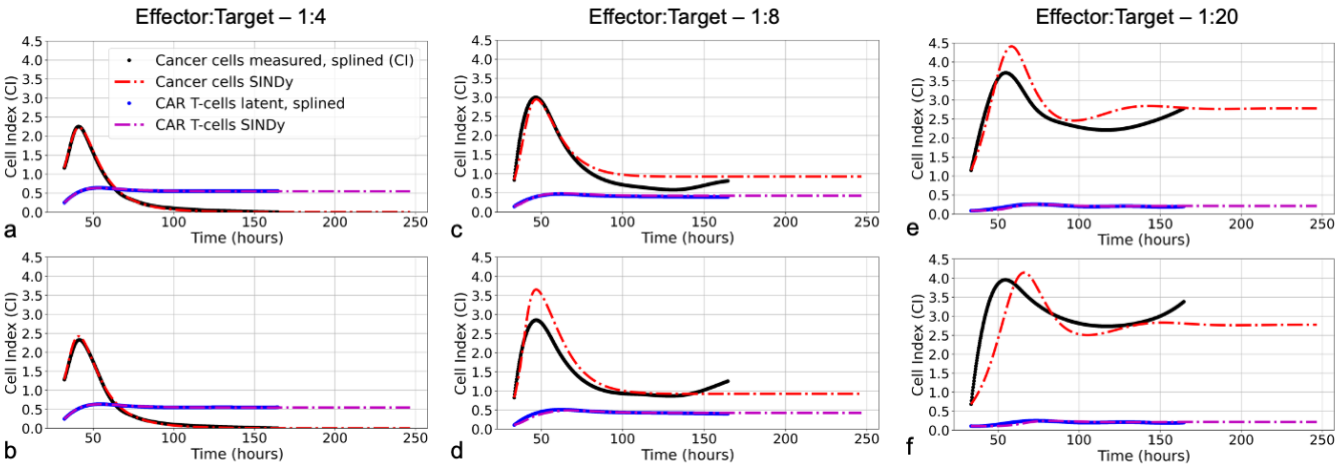


Figure 9. Predicted trajectories of discovered models compared to splined measurements of cancer cells and CAR T-cells. Cancer cell measurements are the black points, while CAR T-cell measurements are the blue points. Predicted trajectories for cancer cells are the red dot-dashed lines, while the CAR T-cells are the purple dot-dashed lines. Model simulations and forward predictions for replicates 1 and 2 for scenarios of (**a-b**) high E:T, (**c-d**) medium E:T, and (**e-f**) low E:T. Note that the best fits between predictions and measurements occur in the high E:T scenario, where assumptions made regarding treatment success and low cancer cell populations in determining model candidate terms are best adhered to. As the E:T ratios get smaller, increasing deviation between discovered model predictions and splined measurements can be qualitatively observed.

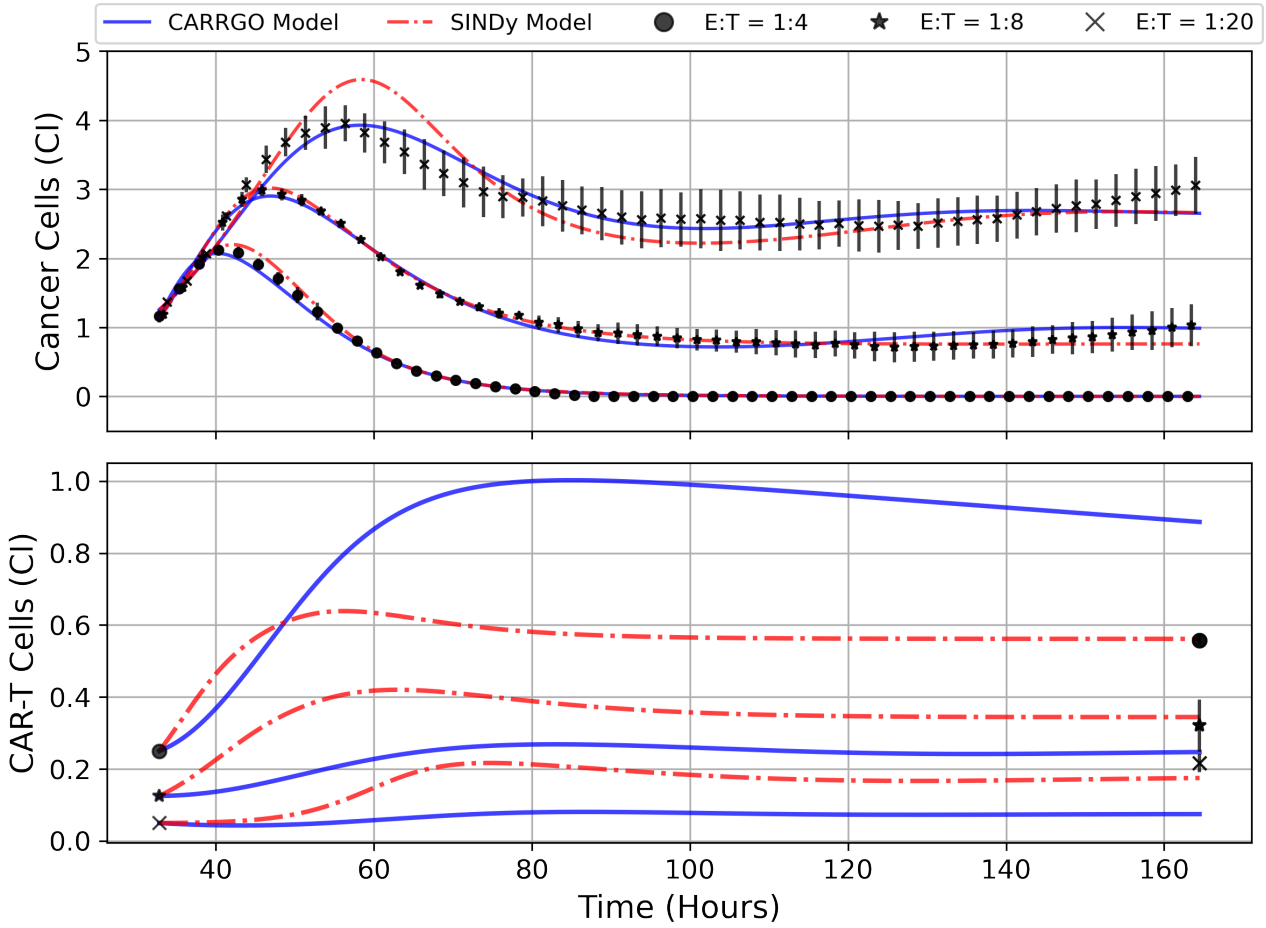


Figure 10. Predictions of cell trajectories for E:T ratios of 1:4, 1:8, and 1:20 from CARRGO model (blue) and SINDy model (red). Model fits were performed using Levenberg-Marquadt Optimization on data aggregated across experimental replicates. Data points represent the mean of all experimental replicates, while error bars represent the ranges across replicates. Of note are the differences in CARRGO and SINDy model predictions for the final CAR T-cell values compared to measurements, and the notable difference in when the maximum CAR T-cell population is reached between CARRGO and SINDy models.

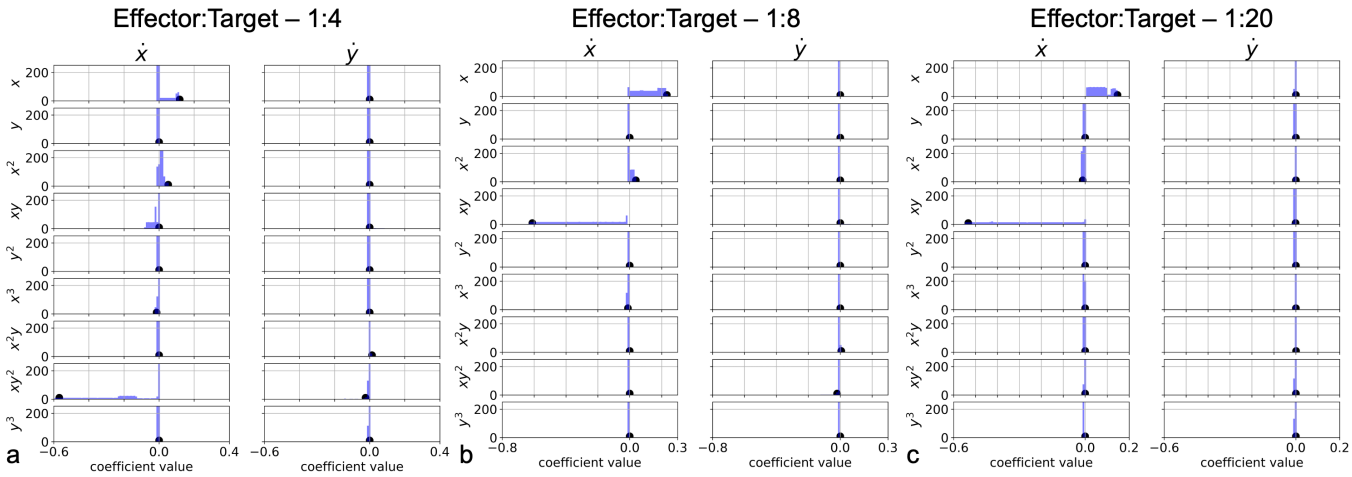


Figure 11. Frequency of discovery for model terms as threshold λ is varied across 1000 values from the interval $[5^{-3}, 10^{1}]$ for **a** high E:T, **b** medium E:T, and **c** low E:T. Black circles indicate values for coefficients corresponding to the selected model based on the Pareto front analysis in Figure 8.

6 TABLES

Table 1. Coefficients for discovered growth model terms across all effector to target ratios.

$E:T$	Growth of cancer cells (x)	Growth rate ρ_x (hrs^{-1})	Carrying capacity K_x (CI)	Allee constants A_x, B_x (CI)	Growth of CAR T-cells (y)	Growth rate ρ_y (hrs^{-1})	Carrying capacity K_y (CI)	Allee constants A_y, B_y (CI)
1 : 4	Weak Allee	0.121	4.792	1.421	Logistic	0.191	0.544	–[1]
1 : 8	Weak Allee	0.237	6.413	3.08	Logistic	0.112	0.313	–
1 : 20	Logistic	0.15	12.5	–	–	–	–	–

[1] – indicates term not discovered.

Table 2. Coefficients for discovered interaction model terms across all effector to target ratios.

$E:T$	Response of CAR T-cells	Type I & II response α (CI^{-1} hrs^{-1})	Type II & III response β (CI^{-2} hrs^{-1})	Cancer cell-CAR T-cell binding	Single binding $\tilde{a}$ (CI^{-1} hrs^{-1})	Double binding $\tilde{b}$ (CI^{-2} hrs^{-1})	Double binding $\tilde{c}$ (CI^{-2} hrs^{-1})
1 : 4	Type II	0.035	−0.009	Double	–[1]	0.593	–
1 : 8	Type II	0.051	−0.009	Single	0.626	–	–
1 : 20	Type III	−0.002	0.005	Mixed	0.545	–	0.063

[1] – indicates term not discovered.

Table 3. Fitting statistics for CARRGO and SINDy models and comparison of shared parameters. Fitting statistics considered are the reduced chi-squared, $\tilde{\chi}^2$, the Akaike information criteria (AIC) and Bayesian information criteria (BIC). Of note are the scores indicating a better fit for the CARRGO model at E:T = 1:4 and 1:20, despite differences in the endpoint CAR T-cell population predictions in Figure 10. Furthermore, we observe generally favorable agreement between parameter estimates, suggesting the *data first* approach of SINDy as a viable alternative to traditional *model first* parameter inference methods.

Model –E:T	$\tilde{\chi}^2$	AIC	BIC	Cancer growth rate ρ (hrs^{-1})	Cancer carrying capacity K_x (CI)	Cancer killing $\tilde{a}$ (CI^{-1} hrs^{-1})	CAR T response α (CI^{-1} hrs^{-1})
CARRGO –1:4	13.6	1380	1400	0.471	3.70	0.555	0.0318
SINDy –1:4	23.0	1660	1700	0.116	4.78	–[1]	0.0327
CARRGO –1:8	0.919	-39.6	-18.3	0.361	6.82	1.26	0.015
SINDy –1:8	0.401	-474	-440	0.190	7.06	0.588	0.0436
CARRGO –1:20	3.14	609	631	0.206	7.69	1.81	0.0195
SINDy –1:20	3.55	674	700	0.123	11.1	0.540	-0.0024

[1] – indicates term not discovered.