Prediction of the Growth Rate Population Dynamic of Bacteria by Causal Jump Dynamic Mode Decomposition

Shara Balakrishnan, Aqib Hasnain, Nibodh Boddupalli, Dennis M. Joshy, and Enoch Yeung

***Abstract*— In this paper, we consider the problem of learning a predictive input-output model of cell growth rate from parametric conditions defining the growth medium. We first introduce a generic data-driven framework for training operator-theoretic models to predict cell growth rate. We then introduce the experimental design and data generated in this study, namely growth curves of *Vibrio natriegens*, as a function of titrated casein and glucose levels. We then evaluate the performance of a classical algorithm for training Koopman operators, the Hankel dynamic mode decomposition algorithm, and show it is unable to predict the biological growth curves. We introduce a modified version of the Hankel dynamic mode decomposition algorithm, that leverages causal-jump embeddings to predict growth rate. We show this algorithm is able to predict growth curves as a function of casein, but that glucose concentration is ultimately non-informative. Our work suggests a method for designing media to achieve optimal growth rates in organisms, whether the goal is to achieve a slow growing microbe or a fast growing one.**

I. Introduction

One of the most fundamental processes in life is the ability to replicate and pass on hereditary material [1]. From viral particles to bacteria to mammalian cells, cell division is fundamental to growth, maintenance of physiological health, and intrinsically tied to the notion of senescence [2].

The mechanisms for controlling growth in organisms are determined by metabolic networks [3], [4], namely their topological structure and parametric realization. Known metabolic networks in well studied model organisms such as *E. coli* [5] and *S. cerevisiae* [6], [7] have given rise to predictive models that translate environmental activity to metabolic network state, and ultimately to predictions of growth rate. For canonical biological model systems, these models have been highly accurate in predicting growth rate and found utility in industrial microbiology applications, e.g. in the design of bioreactors or informing best practices in food safety.

For many biological life forms, relatively little is known about their metabolic network or structure. This is especially the case when developing bioengineering tools in novel host microbes [8], [9]. For new organisms, canonical metabolic networks are lacking and often obtained through a process of sequence alignment and comparative analysis with existing metabolic network models in relative species. However, many novel strains do not exhibit significant similarity, and even in the case of sequence similarity, small mutations can lead to dramatically different growth phenotypes, e.g. growth of non-pathogenic soil strains [10], [11] versus pathogenic counterparts [12]. The absence of predictive cross-species models, as well as the inability to predict growth phenotype wholly from sequence data, motivates the need for data-driven methods to accelerate the discovery of metabolic models and growth rate prediction models.

Due to advances in high-throughput experimental techniques, it is relatively easy to characterize growth rates as a function of exposure to environment. Liquid and acoustic-liquid handling robotics enables interrogation of thousands of growth conditions in a single microtiter plate, which in turn opens the door for using data-driven approaches [13] to predict growth rate as a function of environmental state. Is it possible to accurately predict the growth rate of a microbe, entirely from the chemical composition and environmental parameters of its growth condition? In this paper we explore a data-driven operator theoretic approach that utilizes microtiter plate reader data, and more generally multi-variate time-series data, to develop predictive models of growth rate in *Vibrio natriegens*, one of the fastest growing organisms in the world and a target workhorse [14].

A broadly successful class of data-driven modeling approaches stem from the study of Koopman operators, a mathematical construct for representing the time-evolution of nonlinear dynamical systems. In Koopman operator theory, the time-evolution of a nonlinear system is defined on a function space, acting on the original state of the system. In this function space, known the observables space, the Koopman operator is a linear operator, enabling spectral analysis, the decomposition of eigenspaces, and study of nonlinear structure [15]. The Koopman operator framework has been developed for continuous [16] and discrete time systems [17], [18], for open-loop [17] and input-controlled [19], [20] dynamic systems. Koopman operators generically can be divided into two categories, those with discrete (countable) spectra [17] and those with (uncountable) spectra [15]. Thus, Koopman operators present a powerful framework for analyzing the behavior of nonlinear systems, including predicting or forecasting behavior in a data-driven context.

Many numerical methods for discovering Koopman operators directly from data have been developed in the last two decades [21]–[28]. The most classic approach is the use dynamic mode decomposition (DMD), which models nonlinear dynamics via an approximate local linear expansion [25]. Next there is extended dynamic mode decomposition, which uses an extended dictionary of basis functions [17] or functions with universal function approximation properties to discover an approximation of the lifting map or observables. These techniques suffer from combinatorial explosion, which generally has prohibited analysis of high-dimensional nonlinear systems [29].

The most recent developments in the field of dynamic mode decomposition integrate established advances in deep learning with dynamic mode decomposition [27], [30]–[32]. The promise of these deep learning techniques is that deep neural networks have phenomenal capacity to approximate exponentially many distinct observable functions, at a linear increase in parameter complexity by increasing neural network depth [27]. An additional layer allows for a combinatorial expansion in the number of possible functions that can be expressed. Thus, neural networks-based approaches to deep learning have demonstrated the ability to use low-dimensional dictionaries to discover nonlinear dynamics, on systems that traditionally would require high-dimensional dictionary functions (or were previously intractable) [27]. Recent work has shown deep Koopman learning algorithms can be extended to synthesize controllers for systems subject to uncertainty [33], suggesting that deep Koopman learning can be used broadly for robust controller synthesis.

In this paper, we propose a new dynamic mode decomposition algorithm to predict growth rate as a function of media conditions, resolving the acausality of the Hankel dynamic mode decomposition algorithm with a modified causal-jump operator. In Section II we formulate the Koopman operator and introduce the method of extended dynamic mode decomposition. In Section III we introduce an experimental dataset, the design of the experiment and visualize the data. Section IV reviews the Hankel dynamic mode decomposition algorithm and evaluates its performance, showing it is unable to predict the growth curves with any degree of accuracy in extended forecasting. In Section V we introduce the causal-jump dynamic mode decomposition algorithm, a new algorithm that adapts the idea of delay embeddings and Hankel dynamic mode decomposition, but resolves a long standing issue with causality. We show that the algorithm is able to train a predictive Koopman operator, that predicts with 3.4% on the training data and 9% on the test data on extended forecasting tasks approximately 500 time steps ahead.

II. Koopman Operator Formulation

Consider a discrete-time autonomous nonlinear dynamical system

$$x[k+1] = f(x[k]) \tag{1}$$

with $f : \mathbb{R}^n \to \mathbb{R}^n$ is analytic. Then, there exists a Koopman operator [34] of (1), which acts on a function space $\mathcal{F}$ as $\mathcal{K} : \mathcal{F} \to \mathcal{F}$. This action can be given by

$$\mathcal{K}\psi(x[k]) = \psi \circ f(x[k]). \tag{2}$$

where the function $\psi : \mathbb{R}^n \to \mathbb{R}$ is called an *observable* of the system and the set of all observables $\psi \triangleq \{\psi_i\}_{i=1}^{p}, p \leq \infty$ on the system. Here $\mathcal{F}$ is invariant under the action of $\mathcal{K}$.

The most important property of the Koopman operator that we utilize is the linearity of the operator, in other words,

$$\mathcal{K}(\alpha\psi_1 + \beta\psi_2) = \alpha\psi_1 \circ f + \beta\psi_2 \circ f = \alpha\mathcal{K}\psi_1 + \beta\mathcal{K}\psi_2$$

which follows from (2) since the composition operator is linear. Thus, we have that the Koopman operator of (1) is a linear operator that acts on observable functions $\psi(x_k)$ and propagates them forward in time.

A. DMD and relevant variants

The practical identification of Koopman operator for a nonlinear system from input-output data is commonly done using DMD [25] or extended DMD [17] which constructs an approximate Koopman operator K. Rowley et. al showed that the finite-dimensional approximation to the Koopman operator obtained from DMD is closely related to a spectral analysis of the linear but infinite-dimensional Koopman operator [18]. The approach taken to compute an approximation to the Koopman operator in both DMD and extended DMD is as follows

$$K = \min_{K} ||\Psi(X_f) - K\Psi(X_p)|| = \Psi(X_f)\Psi(X_p)^{\dagger} \tag{3}$$

where $X_f \equiv \begin{bmatrix} x_1 & \dots & x_{N-1} \end{bmatrix}$, $X_p \equiv \begin{bmatrix} x_2 & \dots & x_N \end{bmatrix}$ are snapshot matrices formed from the discrete-time dynamical system (1), $\Psi(X) \equiv \begin{bmatrix} \psi_1(x) & \dots & \psi_R(x) \end{bmatrix}$ is the mapping from physical space into the space of observables and † denotes the Moore-Penrose pseudoinverse. Here N is the number of snapshots i.e. timepoints. We note that DMD is a special case of extended DMD where $\psi(x) = x$. Throughout the rest of the paper, when we refer to the Koopman operator we mean the finite dimensional approximation to the infinite-dimensional Koopman operator.

III. Experimental Setup

To demonstrate the effectiveness of this novel modeling approach, we chose *Vibrio natriegens* for the bacterial specimen for this experiment primarily due to its enormously high growth rate [35]. It also serves as a potential candidate for hosting genetic circuits. The key advantages that this offers to our experiment is the ability to see the growth curve dynamics quickly thereby decreasing the sampling time and enabling rapid proof of concept testing. We formulate the problem of describing *V. nat.*'s growth rate as an input-output problem. The inputs are specific pulse inputs of Casein and Glucose added to *V. nat.*'s culture medium (described below). The output dataset is the rate of growth calculated from the variation of optical density obtained from periodic OD600 measurements from a plate reader.

Incubating *V. nat.*: *V. natriegens* cryo-preserved at -70^oC in 30%(vol/vol) glycerol stock is revived by suspending a small portion into a polypropylene test tube containing 4mL Lysogeny Broth (LB). This is cultured at 37^oC at 200rpm for 4 hours. A cloudy culture medium inside the tube indicates the successful formation of a *V. nat.* culture.

Solution Preparation: The medium chosen for the growth of *V. nat.* is the VN minimal medium (a modified CGXII medium) which was used in [ref] to study the individual effect of various inputs like sugars and intermediates in the metabolic pathway. VN minimal medium includes a combination of salts (refer Table I). We prepared it by adding these required salts (mass per liter shown in Table I) to a half

filled 1L autoclaved bottle with Milli-Q water which is later filled up to 1L and mixed. After the dissolution of salts, the medium is sterilized by filtration. Due to the possibility for these salts to dissociate at high temperatures, autoclaving of the solution was not done.

Glucose (Sugar) and Casein (Proteins) are the two inputs that will be added in varying amounts to the culture to model their effects on the growth curve of *V. nat.* and hence a very concentrated solution of each. 500g/L solution of Glucose is created by mixing 4g of Glucose in 8mL of Milli-Q water and vortexed rigorously for 15-30mins and is taken as the 1x concentration Glucose solution. Since Casein does not dissolve in water, Casein acid hydrolysate is used in its place to bypass the solubility issue. The downside is that Casein hydrolysate does not contain the amino acids Tryptophan and Cystine as they are destroyed during acid hydrolysis. 250g/L solution of Casein acid hydrolysate is created by mixing 2g of it in 8mL of Milli-Q water and is taken as the 1x concentration Casein solution.

Serial dilution setup for VNat culture: To measure the patterns of growth of *Vibrio Natriegens* in the modified VN minimal media (VN minimal media with Casein and Glucose), we periodically measured OD600 values for these microbes in a 96-well plate. Each well of this plate contained $300\mu L$ of modified media - $150\mu L$ of VN minimal mixed with a $150\mu L$ input solution containing both Casein and Glucose. The constituent volumes of Casein and Glucose in this latter $150\mu L$ were calculated based on a 2D serial dilution to study a variety of cases.

Using a fresh 96-well plate, we added $250\mu L$ of casein at a concentration $250g/L$ to well A1. Let this be defined as 1x concentration for casein. By filling the remaining wells in row A with casein at 0.5x concentration, one-dimensional serial dilution can be first performed across the rows A-H. We added $150\mu L$ of MilliQ to each well in these rows prior to this. The serial dilution using $100\mu L$ from row A results in wells of the casein concentration diluted by $2.5\times$ the previous well in the same column. Further, we added $25\mu L$ of both the 1x Casein solution and Glucose of concentration of $500g/L$. This enables us to maintain the concentration of Casein at 0.5x in column 2 and the resulting concentration of Glucose is $67.5g/L$ (1x concentration for Glucose). A second serial dilution is performed across the columns (2-11) using $50\mu L$ from column 2. This results in wells whose concentration is $4\times$ diluted compared to the preceding well in the same row.

We retrieved cultures from the incubator, centrifuged them and discarded the supernatant. To remove residual media, we washed the cell pellet in PBS solution twice using the vortex machine and a centrifuge. To the cell pellet, the VN minimal media is added and vortexed. Finally, we re-suspended the cell pellet in VN Mininal media using the vortex machine. $150\mu L$ of this culture is added to each well of the microplate for a total volume of $300\mu L$ in each well and was used in the plate reader experiment.

Data Collection: The microplate reader is set to 37^oC and the shaker to 807cpm continuous double orbital shaking.

TABLE I: Salts required to prepare 1 litre of VN minimal media

Salt	mass/L	salt	mass/L
Ammonium Sulphate	5g	Iron Sulphate	16.4mg
Sodium Chloride	15g	Manganese Sulphate	10mg
Potassium Phosphate Monobasic	1g	Copper Sulphate	0.3mg
Potassium Phosphate Dibasic	1g	Zinc Sulphate	1mg
Magnesium Sulphate	250mg	Nickel Chloride	0.02mg
Calcium Chloride	10mg	MOPS acid	21mg

Fig. 1: Different initial conditions of substrates obtained by two dimensional serial dilution of Casein and Glucose and the corresponding growth curves are obtained for a period of 48 hours.

The absorbance at 600nm which is termed as the Optical Density (OD600) is measured as a function of time for 48 hours. This serves as the indicator of the cell concentration which when multiplied by the volume of solution gives the number of cells present in each well. Since the volume is maintained constant, cell concentration can be used in the place of cumber of cells. The obtained data along with the inputs are shown in Figure. 1.

IV. Causal-Jump Extended Dynamic Mode Decomposition for Growth Prediction

A. The Growth Rate Modeling Problem

To formulate the problem of optimal input identification for maximal growth rate as a function of time, we need to identify a model that maps the growth rate to the input. The states that represent this system are the bacterial cell count (N_b) and the amount of substrates - Casein (C) and Glucose (G) which change as a function of time (pictorially shown in Figure **??**). The substrates are also the inputs to the system. Hence the general equation of the system is

$$\begin{bmatrix} N_b[k+1] \\ C[k+1] \\ G[k+1] \end{bmatrix} = f(N_b[k], C[k], G[k]) + g(C[k], G[k])$$

where f represents the dynamics and g represent the inputs. In the experimental setup (see section III), the substrates are added only at the initial time and can be treated as an impulse

input. To simplify the scenario, in this paper, the impulse input is treated as an initial condition of the substrates. The system is then simplified in the discrete time framework as

$$\begin{bmatrix} N_b[k+1] \\ C[k+1] \\ G[k+1] \end{bmatrix} = f(N_b[k], C[k], G[k]) \tag{4}$$

given nonzero initial conditions $N_b[0], C[0]$ and $G[0]$. OD600 is the measured quantity in the experiment which has a one-to-one correspondence with N_b.

$$y[k] = OD600[k] = h(N_b) \tag{5}$$

There are identified empirical nonlinear models like that of Monod's [36] which use a single substrate and [37] and [38] which use multiple substrates. Monod's model is a two-state nonlinear dynamical system comprising the substrate(S) and the number of bacteria(N_b):

$$\begin{aligned} \dot{N}_b(t) &= r_{max} \frac{S(t) N_b(t)}{K_s + S(t)} \\ \dot{S} &= -\gamma \dot{N}_b \end{aligned} \tag{6}$$

where r_{max} is the maximum growth rate and K_s is the half velocity constant. In our experimental setup(see section III), the only variable of measurement is OD600 To obtain a model that is more conducive to the measurement framework, we differentiate $\dot{N}_b(t)$ and eliminate the substrate dynamics to obtain

$$\begin{aligned} \ddot{N}_b(t) = \frac{1}{r_{max} K_s N_b} (K_s r_{max} \dot{N_b}^2 + \\ 2\gamma r_{max} N_b \dot{N_b}^2 - \gamma r_{max}^2 N_b^2 \dot{N}_b - \gamma \dot{N_b}^3) \end{aligned} \tag{7}$$

This could potentially serve as the input-output models, but the drawback is that it assumes the growth curve to take a certain shape which may not be necessarily true when culturing new microbes such as *Pseudomonas fluorescens (Pf-5)* or *Vibrio natriegens*(see Figure 1). The consequence is that if we identify the parameters of one model based on a set of initial conditions, it does not necessarily predict the response for other conditions as the behavior could change entirely.

B. The Causal-Jump Extended Dynamic Mode Decomposition Algorithm for Growth Modeling

We instead consider a data-driven approach, using models driven by available measurements and substrate concentrations. In this paper we introduce the causal-jump extended dynamic mode decomposition algorithm, as a variant of the Hankel-DMD algorithm, with the important distinction that the proposed dictionary functions do not violate the property of causality. We briefly review the Hankel DMD algorithm.

1) Hankel Dynamic Mode Decomposition: The Hankel DMD algorithm solves the optimization problem

$$\min_{K, \theta(\Psi)} ||\Psi(X_f) - K\Psi(X_p)|| \tag{8}$$

where

$$\Psi(X_f) = \begin{bmatrix} \psi[t+1](x_0^1) & \dots & \psi[t+1](x_0^{N_T}) \\ \psi[t+2](x_0^1) & \dots & \psi[t+2](x_0^{N_T}) \\ \vdots & \ddots & \vdots \\ \psi[t+n](x_0^1) & \dots & \psi[t+n](x_0^{N_T}) \end{bmatrix}$$

and

$$\Psi(X_p) = \begin{bmatrix} \psi[t](x_0^1) & \dots & \psi[t](x_0^{N_T}) \\ \psi[t+2](x_0^1) & \dots & \psi[t+2](x_0^{N_T}) \\ \vdots & \ddots & \vdots \\ \psi[t+n-1](x_0^1) & \dots & \psi[t+n-1](x_0^{N_T}) \end{bmatrix}$$

and n is the number of timepoints drawn from a given time-series trace $x[t]$, N_T is the number of separate traces, initialized from a distinct initial condition $x_0^j, j = 1, .., N_T$, and

$$\psi[t](x_0) \equiv \psi[t, \tau](x_0) \equiv \begin{bmatrix} x[t]^T & x[t+1]^T & \dots & x[t+\tau]^T \end{bmatrix}^T.$$

Thus for any given lifted snapshot $\psi[t](x_0)$, the Hankel DMD generates a model that is non-causal, as it predicts

$$\psi[t+1, x_0] = K\psi[t, x_0]$$

which implies that predictions for $x[t+1]$ are made using state information from $x[t+\delta]$ where $\delta \geq 1$, which is non-causal.

2) The Causal-Jump Extended Dynamic Mode Decomposition Algorithm: We now propose a modification to the Hankel dynamic mode decomposition algorithm and discuss its implications in terms of closure and function approximation theory an alternative view of snapshots of the dynamical system (1).

To maintain causality, we consider windows of length $\tau \in \mathbb{Z}_{>0}$ but apply a downsampling strategy that ensures the prediction task does not require utilizing future state information to predict the present state. Let $x[0], x[t+1], ..., x[t+N]$ be a state trajectory for the dynamical system (1). Suppose $N+1$ is an integer multiple of τ. Then we divide the state trajectory into jump partitions:

$$\begin{aligned} x[0], x[1], ..., x[\tau - 1] \\ x[\tau], x[\tau + 1], ..., x[2\tau - 1] \\ \vdots \\ x[(R-1)\tau], ..., x[R\tau - 1] \end{aligned} \tag{9}$$

where $R\tau - 1 = N$. We can represent these windows through function composition of the vector field $f(x)$ via the following relation,

$$\begin{bmatrix} (x[(k)\tau) \\ f(x[k\tau]) \\ \vdots \\ f^{(\tau)}([x[k\tau]) \end{bmatrix} = f^\tau \left(\begin{bmatrix} x[(k-1)(\tau))] \\ f(x[(k-1)(\tau)] \\ \vdots \\ f^{(\tau)}(x[(k-1)(\tau)]) \end{bmatrix} \right) \tag{10}$$

We refer to this new structured dynamical system as the τ-jump system of the original dynamical system (1). As long as

f is locally Lipschitz with Lipschitz constant L, it is easy to see via compositional arguments with the Lipschitz property that f^τ will also be locally Lipschitz with Lipschitz constant L^τ and therefore f^τ retains a unique solution for the τ-jump dynamical system. Furthermore, if $f(x)$ is analytic, then the τ-jump system also admits the existence of a Koopman operator

Theorem 1: Let $f(x)$ in system (1) be analytic. Then the τ-jump system also admists a Koopman operator.

Proof: Since $f(x)$ is analytic, it admits a countable-dimension Koopman operator K, with a invariant subspace isomorphic to either a finite or infinite Taylor polynomial basis [34]. Moreover, isomorphism with a Taylor polynomial basis ensures that the Koopman observable space contains the full state observable, i.e. it is state inclusive [**?**]. This implies that the τ step jump from τ function compositions of f can be modeled via the action of the Koopman operator, in the following way:

$$\psi(x[k\tau]) = \psi(f(x(k\tau-1))) = K\psi(x(k\tau-1)) = K \quad (11)$$

where ψ represents the 1-step Koopman observable function.

$$\begin{aligned} x(k\tau) &= f^\tau(x(k-1)\tau) \\ &= f^{(\tau-1)} \circ f((x(k-1)\tau+1) \\ &= f^{(\tau-1)} \circ K_x\psi(x((k-1)\tau)) \\ &= f^{(\tau-2)} \circ K_x\psi(K_x\psi(x((k-1)\tau))) \\ &= g^{(\tau)}(x(k-1)\tau) \end{aligned} \quad (12)$$

where $g(x) = K_x\psi(x)$. . This provides an explicit form for the recovery of the τ-jump governing equations from the one-step Koopman operator and its observable function.

There are two easy arguments to conclude the proof. First, note that since f is analytic, f^τ is analytic and thus by the same reasoning as in [34], f^τ thus must admit a Koopman operator. The second argument is a constructive one, noting that equation

$$\psi(x[(k)\tau]) = K^\tau\psi(x[(k-1)(\tau)) \quad (13)$$

must hold due to τ applications of the 1-step Koopman equation. This means therefore that the following *matrix* equation must hold

$$\psi\left(\begin{bmatrix} x[(k)\tau] \\ x[k\tau+1] \\ \vdots \\ x[(k+1)\tau-1] \end{bmatrix}\right) = \mathbf{K}_J\psi\left(\begin{bmatrix} (x[(k-1)(\tau))] \\ (x[(k-1)]\tau+1]) \\ \vdots \\ (x[(k)\tau-1)] \end{bmatrix}\right) \quad (14)$$

where $\mathbf{K}_J = \text{diag}\left(K^\tau, K^\tau, \dots K^\tau\right)$. This concludes the proof. ∎

The power of the jump partitioning is apparent when considering multiple trajectories. Notice that an element of the stacked data matrix $\Psi(X_p)$ has access to all τ entries of the state vector to define any monomial, polynomial, or higher-order nonlinear observable, without violating the rule of causality, due to the partitioning of the state trajectory. This is precisely the approach we take with causal-jump extended dynamic mode decomposition. The jump-Koopman learning problem we solve is as follows

$$\min_{\mathbf{K_J}} ||\Psi(X_f) - \mathbf{K_J}\Psi(X_p)||$$

$s.t.$

$$\psi\left(\begin{bmatrix} x[(k)\tau] \\ x[k\tau+1] \\ \vdots \\ x[(k+1)\tau-1] \end{bmatrix}\right) = \mathbf{K}_J\psi\left(\begin{bmatrix} (x[(k-1)(\tau))] \\ (x[(k-1)]\tau+1]) \\ \vdots \\ (x[(k)\tau-1)] \end{bmatrix}\right)$$

where

$$\Psi(X_p) = \begin{bmatrix} \psi[0,\tau](x_0^1) & \dots & \psi[0,\tau](x_0^{N_T}) \\ \psi[1,\tau](x_0^1) & \dots & \psi[1,\tau](x_0^{N_T}) \\ \vdots & \ddots & \vdots \\ \psi[R-1,\tau](x_0^1) & \dots & \psi[R-1,\tau](x_0^{N_T}) \end{bmatrix}$$

and

$$\Psi(X_f) = \begin{bmatrix} \psi[1,\tau](x_1^1) & \dots & \psi[1,\tau](x_0^{N_T}) \\ \psi[2,\tau](x_0^1) & \dots & \psi[2,\tau](x_0^{N_T}) \\ \vdots & \ddots & \vdots \\ \psi[R,\tau](x_0^1) & \dots & \psi[R,\tau](x_0^{N_T}) \end{bmatrix}$$

and with a slight abuse of notation (for brevity)

$$\psi[k, x_0] = \psi\left(x(k\tau), x(k\tau+1), ..., x(((k+1)\tau-1)\right) \in \mathbb{R}^{n_L}$$

where n_L is the lifting dimension of the extended basis defined on the set of nonlinear functions on $x(k\tau+1), ..., x((k+1)\tau-1)$. The solution to the causal-jump Koopman learning problem follows that of E-DMD, e.g. using the companion matrix method or singular value decomposition to compute robust estimates of the Moore-Penrose inverse on the data matrices. Further, it can be shown that for analytic functions expressed by universal function approximators with algebraic closure properties on function composition, e.g. logistic, RELU and infinite polynomial bases, these extended causal-jump lifting functions can generate a Koopman invariant subspace. The proof is omitted here due to space constraints.

3) Causal-Jump Extended Dynamic Mode Decomposition for Learning Predictive Growth Models: We now apply our algorithm to generate predictive growth models for cultures of Vibrio natriegens, as a function of environmental parameters such as casein and glucose. The key ideas to take from the Monod model are that the system dynamics can be represented by the number of cells $N_b[k]$ as a function of time step k and that the substrate dynamics can be viewed as an initial condition of the population cell dynamics. This yields the model:

$$N_b[k+1] = f(N_b[k]) \quad (15)$$

starting with the initial conditions $N_b[0], C[0]$ and $G[0]$. The caveat with this model is that the dynamics has no representation of the input and the initial conditions are independent

since the bacteria and substrates are independently added into each well. To bypass this issue, we exploit the fact that we witnessed in (7) that the substrate dynamics is contained in the bacterial cell growth dynamics. Since (7) also uses second derivatives, it is clear that we need to include more past information to estimate the current state. So, we break the whole response across into fragments and represent each state by a collection of time points

$$x[k] = \begin{bmatrix} y[nk] & y[nk+1] & ... & y[nk+n-1] \end{bmatrix}^T \quad (16)$$

where n represents the number of data points in each fragment. Using this representation of state, we can create a model to map the initial condition of the new state to the old one and also predict the cell growth dynamics with more information. The model then translates to

$$\begin{aligned} x[k+1] &= F(x[k]) \\ \begin{bmatrix} N_b[0] \\ C[0] \\ G[0] \end{bmatrix} &= Ax[0] \end{aligned} \quad (17)$$

With the model identification objective in mind, we gear towards system identification techniques of data-based time series models. Since we are dealing with a nonlinear model, we simplify the problem by identifying an approximation of the Koopman operator. The first step towards that is to construct a dictionary of observables by computing all possible monomials of x at each time instant k.

$$\begin{aligned} \psi(x[k]) = [&y[nk], ..., y[nk+n-1], \\ &y^2[nk], y[nk]y[nk+1], ..., y^2[nk+n-1] \\ &y^3[nk], y^2[nk]y[nk+1], ...]^T \end{aligned}$$

Assuming we have N time samples of $x[k]$ and M data-sets, we collect all the observables as

$$\begin{aligned} \Psi_i(x^-) &= \begin{bmatrix} \psi(x_i[0]) & \psi(x_i[1]) & ... & \psi_i(x[N-2]) \end{bmatrix} \\ \Psi_i(x) &= \begin{bmatrix} \psi_i(x[1]) & \psi_i(x[1]) & ... & \psi_i(x[N-1]) \end{bmatrix} \\ & \qquad i = 1, 2, ...M \end{aligned}$$

$$\begin{aligned} \tilde{\Psi}(x^-) &= \begin{bmatrix} \Psi_1(x^-) & \Psi_2(x^-) & ... & \Psi_M(x^-) \end{bmatrix} \\ \tilde{\Psi}(x) &= \begin{bmatrix} \Psi_1(x) & \Psi_2(x) & ... & \Psi_M(x) \end{bmatrix} \end{aligned} \quad (18)$$

Then we can pose the Koopman learning problem as

$$\tilde{\Psi}(x) = K\tilde{\Psi}(x^-) \quad (19)$$

which can be solved by DMD as highlighted in section II. To test the model on a new data set, we take estimate $\psi(x[0])$ and then predict the response using

$$\hat{\psi}(x[k]) = K^k \psi(x[0]) \quad (20)$$

and the required dynamics can be obtained by simply dropping the additional observables other than the base state. The other model to be constructed is the mapping between the initial states. For this purpose, across the M training sets, we collect $\psi(x[0])$ and obtain

$$\Psi(x[0]) = \begin{bmatrix} \psi_1(x[0]) & \psi_2(x[0]) & ... & \psi_M(x[0]) \end{bmatrix}.$$

Since $y_i[0]$ is the first entry of $\psi(x[0])$, it does not require any mapping. So, we construct a matrix of initial substrate conditions.

$$Y_0 = \begin{bmatrix} C_1[0] & C_2[0] & ... & C_M[0] \\ G_1[0] & G_2[0] & ... & G_M[0] \end{bmatrix}.$$

We again formulate the similar structure

$$Y_0 = K_0 \Psi(x[0]) \quad (21)$$

which can again be solved by SVD formulation in Section II.

V. Results

From the data-sets obtained in the plate reader experiments shown in Figure. 1, we identified approximate Koopman operators using Hankel DMD, Extended DMD and the Causal Jump DMD by using the training set of well A7. To ensure equal comparison of the three algorithms, we keep the number of points used to represent the initial condition same in Hankel DMD and CJDMD at 10 points and the highest order of monomials same across EDMD and CJDMD as 10. Using the first 80% of the singular values, we see that EDMD does a good job but CJDMD fits almost perfectly as seen in 2 where HankelDMD fails miserably.

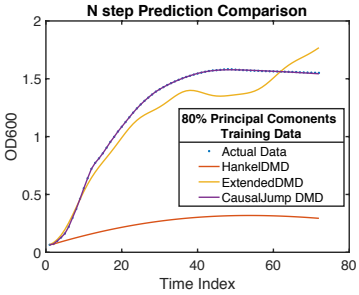

Fig. 2: Comparing the goodness of fit of various DMD algorithms with a single training set with 80% of the singular values

The N step prediction capability of these models were tested in a fresh dataset of well A8 and the results are seen in Figure. 3. CJDMD fit with a absolute mean percent error

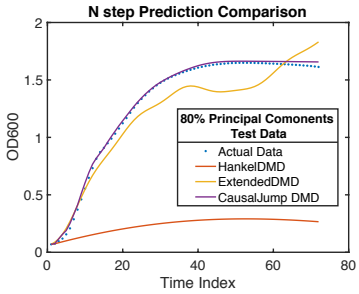

Fig. 3: Comparing the goodness of fit of the various DMD algorithms based on a test set. It can be seen that Causal Jump DMD does a splendid job in both test set and training set with 80% of the singular values

of 1.5%. We then took the wells A5 , A7, A8, A9, A11, B1, B6, B8, B9, B10, C1, C6, C7, C8 and C9 as training

sets and A6, A10, B7, B11 C10 and C11 as the test sets to build a better model and predict across many variations. The remaining data sets exhibit different trends and are not considered.

In the training process, the two parameters that we can tweak to optimize the model are n in equation (16) and maximum order of monomials in (18). By choosing $n = 10$ and keeping the maximum order or monomial to be 10, the Koopman operator has been identified and the prediction on the training data is shown in Figure. 4 and on the test set is shown in Figure. 5.

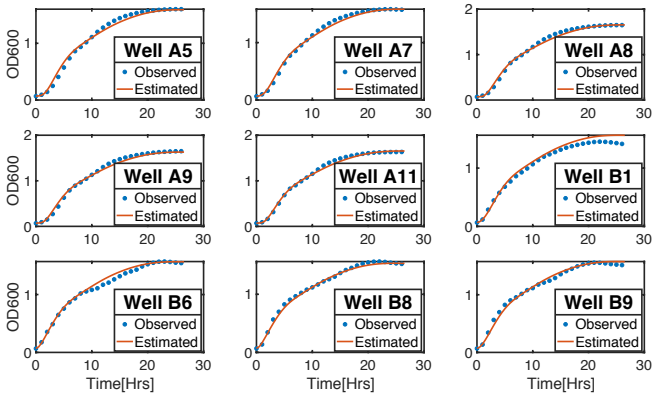

Fig. 4: The identified Koopman operator is tested on the training sets with 10 point initial condition and up to 10^{th} order monomials to get a MSE of 3.4%

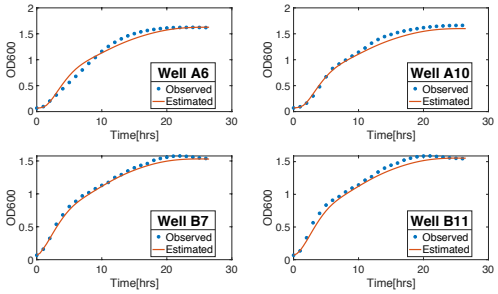

Fig. 5: The identified Koopman operator is tested on the test sets by using the initial observables$\psi(x[0])$ and the mean squared error remains the same as that of the training set

Both show very similar Mean Square error of 3.4% on the training data and 9% on the test data ensuring the goodness of the model. A model for mapping $\psi(x[0])$ to $C[0]$ and $G[0]$ has been identified using (21) and tested on both training and test data and shown in Figure. 6. The Casein model performs very well on the test data as well with less than 5% on training data and less than 15% error on test data. But the glucose model fails miserably in the training data.

This indicates that Glucose does not have a major role to play in the initial dynamics of the cell growth dynamic. This

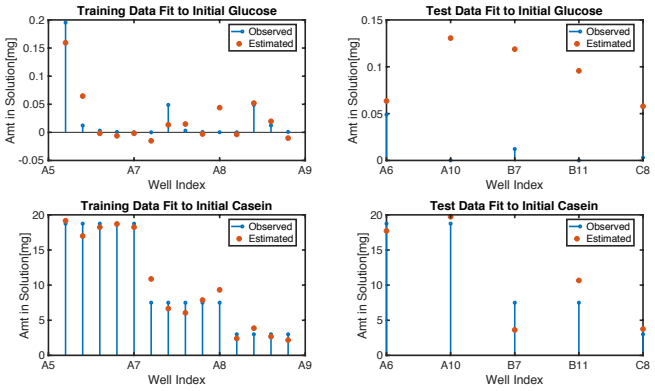

Fig. 6: The initial substrate conditions are mapped to the initialdynamics. The figure indicates that Casein can be predicted with ahigh accuracy reinforcing that the dynamics of Casein is embeddedin the cell growth dynamics. Glucose cannot be predicted properlyneither in the training set nor the test set.

can also be witnessed in the data shown in Figure. 1 that the high growth rate is witnessed only in the region of high Casein and low Glucose and since our data set comprises of only the large growth rate region, the role of Glucose should be minimal.

VI. Conclusion

We have obtained the data of bacterial cell population in response to various initial substrate conditions and developed models that utilize the initial response of the cell culture to map the dynamics at any point in time and also map the initial conditions to generate the response. In the models developed above, the autocorrelation of the prediction error is nonzero at nonzero lags indicating that the Koopman operator has developed the best linear model similar to an Output-Error model in linear system identification. The future scope of the work involves modeling the noise dynamics by stochastic models, remove the simplification of treating impulse input as an initial state and formulate the input-output dynamics, formulating an optimization problem which identifies the best initial conditions to maximize the growth rate. This will pave the way to study the bacteria in unknown environments, robustness to changes in environment and genetic modification of other bacterial species to achieve maximum growth rate which is a boon to Industrial Biology.

References

[1] D. Kornberg and D. TA, "Replication," *San Francisco: W H. Freeman*, 1980.

[2] N. F. Mathon and A. C. Lloyd, "Cell senescence and cancer," *Nature Reviews Cancer*, vol. 1, no. 3, p. 203, 2001.

[3] G. Wu, Q. Yan, J. A. Jones, Y. J. Tang, S. S. Fong, and M. A. Koffas, "Metabolic burden: cornerstones in synthetic biology and metabolic engineering applications," *Trends in biotechnology*, vol. 34, no. 8, pp. 652–664, 2016.

[4] D. S. Glazier, "Is metabolic rate a universal pacemakerfor biological processes?" *Biological Reviews*, vol. 90, no. 2, pp. 377–407, 2015.

[5] D. De Martino, F. Capuani, and A. De Martino, "Growth against entropy in bacterial metabolism: the phenotypic trade-off behind empirical growth rate distributions in e. coli," *Physical biology*, vol. 13, no. 3, p. 036005, 2016.

[6] B. J. Sanchez, C. Zhang, A. Nilsson, P.-J. Lahtvee, E. J. Kerkhoven, and J. Nielsen, "Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints," *Molecular systems biology*, vol. 13, no. 8, 2017.

[7] M. Zwietering, I. Jongenburger, F. Rombouts, and K. Van't Riet, "Modeling of the bacterial growth curve," *Appl. Environ. Microbiol.*, vol. 56, no. 6, pp. 1875–1881, 1990.

[8] T. Tschirhart, V. Shukla, E. E. Kelly, Z. Schultzhaus, E. NewRingeisen, J. S. Erickson, Z. Wang, W. garcia, E. Curl, R. G. Egbert *et al.*, "Synthetic biology tools for the fast-growing marine bacterium vibrio natriegens," *ACS synthetic biology*, 2019.

[9] N. Khan, E. Yeung, Y. Farris, S. J. Fansler, and H. C. Bernstein, "A broad-host-range event detector: expanding and quantifying performance across bacterial species," *bioRxiv*, p. 369967, 2018.

[10] C. Gill and K. Tan, "Effect of carbon dioxide on growth of pseudomonas fluorescens." *Appl. Environ. Microbiol.*, vol. 38, no. 2, pp. 237–240, 1979.

[11] D. M. Gulliver, G. V. Lowry, and K. B. Gregory, "Comparative study of effects of co2 concentration and ph on microbial communities from a saline aquifer, a depleted oil reservoir, and a freshwater aquifer," *Environmental Engineering Science*, vol. 33, no. 10, pp. 806–816, 2016.

[12] A. E. LaBauve and M. J. Wargo, "Growth and laboratory maintenance of pseudomonas aeruginosa," *Current protocols in microbiology*, vol. 25, no. 1, pp. 6E–1, 2012.

[13] A. P. Palacios, J. M. Marín, E. J. Quinto, M. P. Wiper *et al.*, "Bayesian modeling of bacterial growth for multiple populations," *The Annals of Applied Statistics*, vol. 8, no. 3, pp. 1516–1537, 2014.

[14] H. H. Lee, N. Ostrov, B. G. Wong, M. A. Gold, A. Khalil, and G. M. Church, "Vibrio natriegens, a new genomic powerhouse," *bioRxiv*, p. 058487, 2016.

[15] I. Mezic, "Spectral properties of dynamical systems, model reduction and decompositions," *Nonlinear Dynamics*, vol. 41, no. 1-3, pp. 309–325, 2005.

[16] M. Budišić, R. Mohr, and I. Mezić, "Applied koopmanism," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 22, no. 4, p. 047510, 2012.

[17] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, "A data–driven approximation of the koopman operator: Extending dynamic mode decomposition," *Journal of Nonlinear Science*, vol. 25, no. 6, pp. 1307–1346, 2015.

[18] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, "Spectral analysis of nonlinear flows," *Journal of fluid mechanics*, vol. 641, pp. 115–127, 2009.

[19] J. L. Proctor, S. L. Brunton, and J. N. Kutz, "Dynamic mode decomposition with control," *SIAM Journal on Applied Dynamical Systems*, vol. 15, no. 1, pp. 142–161, 2016.

[20] M. O. Williams, M. S. Hemati, S. T. Dawson, I. G. Kevrekidis, and C. W. Rowley, "Extending data-driven koopman analysis to actuated systems," *IFAC-PapersOnLine*, vol. 49, no. 18, pp. 704–709, 2016.

[21] T. Askham and J. N. Kutz, "Variable projection methods for an optimized dynamic mode decomposition," *SIAM Journal on Applied Dynamical Systems*, vol. 17, no. 1, pp. 380–416, 2018.

[22] Y. Kaneko, S. Muramatsu, H. Yasuda, K. Hayasaka, Y. Otake, S. Ono, and M. Yukawa, "Convolutional-sparse-coded dynamic mode decomposition and its application to river state estimation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1872–1876.

[23] O. Azencot, W. Yin, and A. Bertozzi, "Consistent dynamic mode decomposition," *arXiv preprint arXiv:1905.09736*, 2019.

[24] K. Manohar, E. Kaiser, S. L. Brunton, and J. N. Kutz, "Optimized sampling for multiscale dynamics," *Multiscale Modeling & Simulation*, vol. 17, no. 1, pp. 117–136, 2019.

[25] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of fluid mechanics*, vol. 656, pp. 5–28, 2010.

[26] S. Sinha and E. Yeung, "On computation of koopman operator from sparse data," *arXiv:1901.03024*, 2019.

[27] E. Yeung, S. Kundu, and N. Hodas, "Learning deep neural network representations for koopman operators of nonlinear dynamical systems," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 4832–4839.

[28] A. Hasnain, S. Sinha, Y. Dorfan, A. E. Borujeni, Y. Park, P. Maschhoff, U. Saxena, J. Urrutia, N. Gaffney, D. Becker *et al.*, "A data-driven method for quantifying the impact of a genetic circuit on its host," *arXiv preprint arXiv:1909.06455*, 2019.

[29] C. A. Johnson and E. Yeung, "A class of logistic functions for approximating state-inclusive koopman operators," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 4803–4810.

[30] S. E. Otto and C. W. Rowley, "Linearly recurrent autoencoder networks for learning dynamics," *SIAM Journal on Applied Dynamical Systems*, vol. 18, no. 1, pp. 558–593, 2019.

[31] N. Takeishi, Y. Kawahara, and T. Yairi, "Learning koopman invariant subspaces for dynamic mode decomposition," in *Advances in Neural Information Processing Systems*, 2017, pp. 1130–1140.

[32] Q. Li, F. Dietrich, E. M. Bollt, and I. G. Kevrekidis, "Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 10, p. 103111, 2017.

[33] P. You, J. Pang, and E. Yeung, "Deep koopman controller synthesis for cyber-resilient market-based frequency regulation," *IFAC-PapersOnLine*, vol. 51, no. 28, pp. 720–725, 2018.

[34] E. Yeung, Z. Liu, and N. O. Hodas, "A koopman operator approach for computing and balancing gramians for discrete time nonlinear systems," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 337–344.

[35] E. Hoffart, S. Grenz, J. Lange, R. Nitschel, F. Müller, A. Schwentner, A. Feith, M. Lenfers-Lücker, R. Takors, and B. Blombach, "High substrate uptake rates empower vibrio natriegens as production host for industrial biotechnology," *Appl. Environ. Microbiol.*, vol. 83, no. 22, pp. e01 614–17, 2017.

[36] J. Monod, "The growth of bacterial cultures," *Annual review of microbiology*, vol. 3, no. 1, pp. 371–394, 1949.

[37] B. W. Brandt, I. M. van Leeuwen, and S. A. Kooijman, "A general model for multiple substrate biodegradation. application to co-metabolism of structurally non-analogous compounds," *Water research*, vol. 37, no. 20, pp. 4843–4854, 2003.

[38] D. S. Kompala, D. Ramkrishna, N. B. Jansen, and G. T. Tsao, "Investigation of bacterial growth on mixed substrates: experimental evaluation of cybernetic models," *Biotechnology and Bioengineering*, vol. 28, no. 7, pp. 1044–1055, 1986.