Identifiability of linear noise approximation models of chemical reaction networks from stationary distributions

Theodore W. Grunberg and Domitilla Del Vecchio

***Abstract*— Biomolecular systems can often be modeled by chemical reaction networks with unknown parameters. In many cases, the available data is constituted of samples from the stationary distribution, wherein each sample is given by a cell in a population. In this work, we develop a framework to assess identifiability of parameters in such a situation. Working with the Linear Noise Approximation (LNA) we give an algebraic formulation of identifiability and use it to certify identifiability with Hilbert's Nullstellensatz. We include applications to particular biomolecular systems, focusing on the identifiability of a sequestration-based motif and of a feedback arrangement based on it.**

I. INTRODUCTION

Identifiability is the property of a system that guarantees that the parameters can be determined from measured data. The strongest form of identifiability is *a priori global structural* identifiability, which ensures that no matter the true values of the parameters, it is possible to estimate them arbitrarily accurately with infinitely many data [5], [19]. For systems modeled by ordinary differential equations (ODEs) with the measured data being the entire trajectory of the output, global a priori identifiability can be assessed via a variety of methods such as those based on Ritt's algorithm [3], [6], [19], or observability analysis whereby the parameters are treated as additional states, which are constant with respect to time [9], [25], [26], [30].

The systems we consider in this paper are those of biomolecular reactions inside the cell. In this case, each cell contains a copy of the chemical reaction network, and thus, in a population of cells, the average across the population at a particular instant in time is the average across many sample paths of the underlying Markov process. If the measurement technique provides data that is the average across the population of cells as a function of time, such as with a plate reader, the chemical reaction network can be modeled by ODEs describing the time evolution of the mean concentrations of a set of molecular species. In this case, identifiability analysis can be carried out using the methods discussed above. However, alternative measurement techniques that do not average over the population exist, such as with flow cytometry. Many of these techniques do not track individual cells between measurement times, but instead measure the population distribution as a function of time. In this case, it is not known in general how much information about the parameters is contained in the data. However, it is known that in certain cases the population distribution gives more information about the parameters than just the means [18], [21], [22], [27]. In general, assessing identifiability in this setting remains an open problem, with Cinquemani giving a method for determining local identifiability of chemical reaction networks where all the reactions have propensities that are affine in the state [10].

An even more restricted variation on the above setting is the case in which instead of measuring samples from the population distribution as a function of time, only samples from the stationary distribution are available. Such a case is of practical importance, since it is often easier to design an experiment where a population of cells is allowed to grow to steady state, at which point the concentrations of the molecular species within each of the cells is measured, such as with single-cell RNA-sequencing [20]. Algorithms for identifying the parameters from such data have been proposed [4], [13], [23], but the question of identifiability has not been addressed, and therefore the proposed methods are not guaranteed to give accurate estimates of the parameters. In this work, we consider the identifiability of chemical reaction networks from stationary distributions. We specifically model chemical reaction networks using the Linear Noise Approximation (LNA) and we algebraically characterize global identifiability from the stationary distribution. Using this algebraic characterization we compute certificates of identifiability based on Hilbert's Nullstellensatz.

This paper is organized as follows. In Section II, we describe the LNA and Hilbert's Nullstellensatz. In Section III, we derive our algebraic characterization of identifiability, and in Section IV we apply our method to certify global identifiability of several chemical reaction networks from stationary distributions.

II. MATHEMATICAL BACKGROUND

A. The linear noise approximation

A chemical reaction network is a system of one or more distinct chemical species, which interact through *reactions*, events that instantaneously change the number of molecules of each species. The LNA model of a chemical reaction network makes the approximation that $\boldsymbol{X} = \Omega\boldsymbol{x} + \sqrt{\Omega}\boldsymbol{\xi}$, where $\boldsymbol{X}$ is the vector of molecular counts of the species, Ω is the volume in which the reactions occur, $\boldsymbol{x}$ is the vector of mean concentrations of the species, and $\boldsymbol{\xi}$ is the stochastic fluctuation of the concentrations about $\boldsymbol{x}$. The LNA is accurate when the molecular counts and the volume

This work was supported in part by the National Science Foundation under CMMI grant 1727189.

T. W. Grunberg is with the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139, USA grunberg@mit.edu

D. Del Vecchio is with the Department of Mechanical Engineering, MIT, Cambridge, MA 02139, USA ddv@mit.edu

are both large [17], [29]. A complete derivation of the LNA can be found in [29].

A chemical reaction network with n species has a state $\boldsymbol{X} = \begin{bmatrix} X_1 & X_2 & \cdots X_n \end{bmatrix}^T$, where X_j is the molecular counts of species j. Each of the r reactions is of the form $\boldsymbol{s}_{ri}^T \boldsymbol{X} \xrightarrow{k_i} \boldsymbol{s}_{pi}^T \boldsymbol{X}$, where $\boldsymbol{s}_{ri}$ is the vector of molecular counts of each species consumed by reaction i, $\boldsymbol{s}_{pi}$ is the vector of molecular counts of each species created when reaction i occurs, and k_i is the reaction rate constant of reaction i. We define

$$\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{k}) = S\boldsymbol{q}(\boldsymbol{x}; \boldsymbol{k}) \tag{1}$$

where

$$S = \begin{bmatrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \cdots & \boldsymbol{s}_r \end{bmatrix},$$

with $\boldsymbol{s}_i = \boldsymbol{s}_{pi} - \boldsymbol{s}_{ri}$. Additionally, $\boldsymbol{k} = \begin{bmatrix} k_1 & \cdots & k_r \end{bmatrix}$ and $\boldsymbol{q}(\boldsymbol{x}; \boldsymbol{k}) = \begin{bmatrix} q_1(\boldsymbol{x}; k_1) & q_2(\boldsymbol{x}; k_2) & \cdots & q_r(\boldsymbol{x}; k_r) \end{bmatrix}^T$. Here $q_i(\boldsymbol{x}; k_i) = k_i \prod_{j=1}^n x_j^{s_{ri}^j}$ is the macroscopic propensity of reaction i, where s_{ri}^j denotes the j^{th} element of $\boldsymbol{s}_{ri}$. The LNA model is then

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t); \boldsymbol{k}),\ \boldsymbol{x}_0(0) = \boldsymbol{x}_0, \tag{2a}$$

$$\dot{\boldsymbol{\xi}}(t) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \boldsymbol{\xi}(t) + \Gamma(\boldsymbol{x}(t); \boldsymbol{k}) \boldsymbol{w}(t),\ \boldsymbol{\xi}(0) = \boldsymbol{\xi}_0, \tag{2b}$$

where $\boldsymbol{w}(t)$ is white, zero mean Gaussian noise with identity covariance, and

$$\Gamma(\boldsymbol{x}; \boldsymbol{k}) = S \operatorname{diag}\left(\sqrt{\boldsymbol{q}(\boldsymbol{x}; \boldsymbol{k})} \right). \tag{3}$$

We assume that (2a) has a unique, globally asymptotically stable equilibrium point, $\boldsymbol{x}^*$, in which case the stationary distribution of $\boldsymbol{X}(t)/\Omega$ under the LNA is $\mathcal{N}(\boldsymbol{x}^*(\boldsymbol{k}), \frac{1}{\Omega} P^*(\boldsymbol{k}))$, where $\boldsymbol{x}^*$ and P^* are the solution to

$$0 = \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{k}), \tag{4a}$$

$$0 = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} P + P \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}^T - \Gamma(\boldsymbol{x}; \boldsymbol{k}) \Gamma(\boldsymbol{x}; \boldsymbol{k})^T. \tag{4b}$$

In the sequel, we attach to a chemical reaction network a function $\mathcal{R}$ that maps reaction rate constants to the stationary distribution given by (4), i.e., $\mathcal{R} : \mathbb{R}^r_{>0} \to \mathbb{R}^n \times \mathbb{S}^{n \times n}$, defined by $\mathcal{R}(\boldsymbol{k}) = \left(\boldsymbol{x}^*(\boldsymbol{k}), \frac{1}{\Omega} P^*(\boldsymbol{k}) \right)$. Here $\mathbb{S}^{n \times n}$ is the space of symmetric $n \times n$ real matrices.

To illustrate the theory we use the following running example throughout this paper.

Example 1 (Illustrative Example 1): Consider a chemical reaction network with one species ($n = 1$) and three reactions ($r = 3$) given by

$$\emptyset \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{X}_1 \xleftarrow{k_3} 2\mathrm{X}_1, \tag{5}$$

where each arrow represents one reaction and reaction i is labeled by k_i, its reaction rate constant. In this case, (1) takes the form

$$\frac{d}{dt} x_1 = \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{k}) = k_1 - k_2 x_1 - k_3 x_1^2. \tag{6}$$

For all $\boldsymbol{k} > 0$ there is a unique and globally asymptotically stable equilibrium point in $\mathbb{R}_{\geq 0}$. Therefore, the LNA gives a unique stationary distribution for all $\boldsymbol{k} > 0$. For this system, $\boldsymbol{q}(\boldsymbol{x}; \boldsymbol{k}) = \begin{bmatrix} k_1 & k_2 x_1 & k_3 x_1^2 \end{bmatrix}^T$ and $S = \begin{bmatrix} 1 & -1 & -1 \end{bmatrix}$. Therefore, from (3) we obtain that

$$\Gamma(x; \boldsymbol{k}) \Gamma(x; \boldsymbol{k})^T = k_1 + k_2 x_1 + k_3 x_1^2. \tag{7}$$

B. Nullstellensatz

In this section, we give an overview of the algebraic geometry tool we use in this work, Hilbert's Nullstellensatz [11]. Hilbert's Nullstellensatz provides a computational method to determine if there are any solutions to a set of polynomial equations. Given $\boldsymbol{z}$, an n'-dimensional vector of variables, we let $\mathbb{Q}[\boldsymbol{z}]$ denote the set of all polynomials in $\boldsymbol{z}$ with rational coefficients. For $p \in \mathbb{Q}[\boldsymbol{z}]$, we let $p(\boldsymbol{z}')$ denote p evaluated at $\boldsymbol{z}' \in \mathbb{C}^{n'}$. Given $\mathcal{P} = \{p_1, p_2, \ldots, p_m\} \subseteq \mathbb{Q}[\boldsymbol{z}]$, the ideal $\langle \mathcal{P} \rangle \subseteq \mathbb{Q}[\boldsymbol{z}]$ generated by $\mathcal{P}$ is defined as

$$\langle \mathcal{P} \rangle = \{ g \in \mathbb{Q}[\boldsymbol{z}] | g = \sum_{i=1}^m \lambda_i p_i,\ p_1, \ldots, p_m \in \mathcal{P}, \\ \lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{Q}[\boldsymbol{z}] \}.$$

We now give a statement of Hilbert's Nullstellensatz.

Theorem 2.1 (Ch. 4 of [11]): Let $\mathcal{P} = \{p_1, p_2, \ldots, p_m\} \subseteq \mathbb{Q}[\boldsymbol{z}]$. Then

$$\emptyset = \left\{ \boldsymbol{z} \in \mathbb{C}^{n'} \middle| 0 = p_1(\boldsymbol{x}), 0 = p_2(\boldsymbol{x}), \ldots, 0 = p_m(\boldsymbol{x}) \right\}$$

if and only if $-1 \in \langle \mathcal{P} \rangle$.

Given a set of polynomials, $\mathcal{P}$, there are multiple ways to check if $-1 \in \langle \mathcal{P} \rangle$. For example, for a fixed maximum degree of the λ_i's, checking if there exist λ_i's such that $-1 = \sum_{i=1}^m \lambda_i p_i$ is equivalent to solving a system of linear equations. However, an alternative approach, which we use in the examples section of this paper, is based on *reduced Gröbner bases* [28], a special set of polynomials associated with $\langle \mathcal{P} \rangle$, which can be computed algorithmically and reveal whether or not $-1 \in \langle \mathcal{P} \rangle$. For completeness, we give a brief summary of the theory of Gröbner bases in Appendix A.

III. MAIN RESULT

We now present the main result of this work, which is an algebraic characterization of global indentifiability of LNA models of chemical reaction networks from stationary distributions. The system identification problem we study is the case where a large number of samples from the stationary distribution of the LNA model of a chemical reaction network are measured, and the goal is to estimate the reaction rate constants $\boldsymbol{k}$. We investigate the simplest case, where infinitely many samples are available, so that the stationary distribution can be reconstructed exactly, and the volume Ω is a known constant. In this case $\boldsymbol{x}^*$ and $\frac{1}{\Omega} P^*$ can be exactly determined from the data. In this case the natural notion of global identifiability for chemical reaction networks from stationary distributions is as follows.

Definition 3.1: A chemical reaction network $\mathcal{R}(\boldsymbol{k})$ is *stationary globally identifiable* over $K \subseteq \mathbb{R}^r_{>0}$ if for any $\boldsymbol{k}_1, \boldsymbol{k}_2 \in K$ such that $\mathcal{R}(\boldsymbol{k}_1) = \mathcal{R}(\boldsymbol{k}_2)$, there exists $a \in \mathbb{R}$ such that $\boldsymbol{k}_2 = a \boldsymbol{k}_1$.

Our definition of identifiability allows for the possibility that all $\boldsymbol{k} \in \operatorname{span} \boldsymbol{k}'$ result in the same stationary distribution. In fact, for any chemical reaction network $\mathcal{R}$, it is true that $\mathcal{R}(\boldsymbol{k}) = \mathcal{R}(\alpha \boldsymbol{k})$ for all $\boldsymbol{k} \in K$ and $\alpha > 0$. Therefore, when only the stationary distribution is measured we can only hope to identify $\boldsymbol{k}$ up to a scaling.

To begin, we show that (4) is a linear equation for $\boldsymbol{k}$, as formalized in the following proposition.

Proposition 3.1: Given a chemical reaction network, there exists a matrix $A(\boldsymbol{x}, P) \in \mathbb{R}^{\frac{n^2+n}{2} \times r}$ such that (4) can be written in the form

$$0 = A(\boldsymbol{x}, P)\boldsymbol{k}, \tag{8}$$

where the entries of $A(\boldsymbol{x}, P)$ are polynomials in the entries of $\boldsymbol{x}$ and in the entries of P on and above the diagonal.

Proof: To begin, we write (4a) as

$$\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{k}) = \sum_{i=1}^{r} k_i \boldsymbol{s}_i \prod_{j=1}^{n} x_j^{s_{ri}^j}$$

and (4b) as $0 = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} P + P \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}^T - S \operatorname{diag} \boldsymbol{q}(\boldsymbol{x}; \boldsymbol{k}) S^T$, where we have used (3) and the fact that $\boldsymbol{q}(\boldsymbol{x}; \boldsymbol{k}) \geq 0$ for all $\boldsymbol{k} \geq 0$. Since $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\boldsymbol{q}(\boldsymbol{x}; \boldsymbol{k})$ are linear in $\boldsymbol{k}$, the right hand side of (4) is linear in $\boldsymbol{k}$ and in P. To complete the proof, observe that (4) has $n + n^2$ equations for $\boldsymbol{x} \in \mathbb{R}^n_{\geq 0}$ and $P \in \mathbb{S}^{n \times n}$. Since P is symmetric, there are only $\frac{n^2+n}{2}$ unique equations in (4b). Therefore we can form $A(\boldsymbol{x}, P) \in \mathbb{R}^{\frac{n^2+n}{2} \times r}$ by removing the repeated equations. ∎

Since P is symmetric, we think of $A(\boldsymbol{x}, P)$ as a function of $\boldsymbol{x}$ and of the $\frac{n^2+n}{2}$ entries of P that are on or above the diagonal.

Example 1 (Illustrative example 1 continued): We want to determine if $\mathcal{R}_1$, given by (5), is stationary globally identifiable over $\mathbb{R}^3_{>0}$. For this chemical reaction network, writing out (4) explicitly yields

$$0 = k_1 - k_2 x_1 - k_3 x_1^2, \tag{9a}$$
$$0 = 2(-k_2 - 2k_3 x_1) p_{11} + k_1 + k_2 x_1 + k_3 x_1^2, \tag{9b}$$

where we have used that $\boldsymbol{x} = x_1$ and $P = p_{11}$ and used (6) and (7). We therefore have that for $\mathcal{R}_1$

$$A(\boldsymbol{x}, P) = \begin{bmatrix} 1 & -x_1 & -x_1^2 \\ 1 & x_1 - 2p_{11} & x_1^2 - 4p_{11}x_1 \end{bmatrix}.$$

Proving that a given system is stationary globally identifiable requires proving that (8) has only one subspace of solutions in $\boldsymbol{k}$ for all $(\boldsymbol{x}, P)$ such that there exists $\boldsymbol{k} \in K$ satisfying $(\boldsymbol{x}, P) = \mathcal{R}(\boldsymbol{k})$. To do this we propose a method based on algebraic geometry. To this end we define the set

$$V = \big\{(\boldsymbol{x}, P, \boldsymbol{k}) \in (\mathbb{R}^n, \mathbb{S}^{n \times n}, \mathbb{R}^r_{>0}) \big| \\ 0 = A(\boldsymbol{x}, P)\boldsymbol{k},\ \operatorname{rank}(A(\boldsymbol{x}, P)) < r - 1\big\}. \tag{10}$$

Using V we now give an algebraic condition for a chemical reaction network to be stationary globally identifiable.

Theorem 3.1: A chemical reaction network $\mathcal{R}$ is stationary globally identifiable over $\mathbb{R}^r_{>0}$ if and only if $V = \emptyset$.

Proof: First, suppose $\mathcal{R}$ is not stationary globally identifiable over $\mathbb{R}^r_{>0}$. Then there exists $\boldsymbol{k}_1, \boldsymbol{k}_2 > 0$, with $\boldsymbol{k}_2$ and $\boldsymbol{k}_1$ linearly independent, such that $0 = A(\boldsymbol{x}, P)\boldsymbol{k}_1$ and $0 = A(\boldsymbol{x}, P)\boldsymbol{k}_2$. This immediately implies that $\operatorname{rank} A(\boldsymbol{x}, P) < r - 1$, and therefore $(\boldsymbol{x}, P, \boldsymbol{k}_1) \in V$. Now suppose that there exists $(\boldsymbol{x}', P', \boldsymbol{k}') \in V$. By the definition of V, $\operatorname{rank} A(\boldsymbol{x}', P') < r - 1$, and so there exists a W, a subspace of dimension 2 containing $\boldsymbol{k}$ such that $0 = A(\boldsymbol{x}', P')W$. It then follows from the fact that $\mathbb{R}^r_{>0}$ is open that there exists $\boldsymbol{k}'' > 0$, linearly independent from $\boldsymbol{k}'$, such that $0 = A(\boldsymbol{x}, P)\boldsymbol{k}''$ and therefore $\mathcal{R}$ is not stationary globally identifiable. ∎

Given a chemical reaction network $\mathcal{R}$, from Theorem 3.1 we have a set defined in terms of polynomial equations and inequalities, which if empty proves that $\mathcal{R}$ is globally stationary identifiable. To transform this condition into one that uses a set defined with only polynomial equalities, we require the following result, which is a consequence of the standard determinant characterization of rank, which we now state.

Lemma 3.1: (Determinant rank characterization) Let $A \in \mathbb{R}^{n \times m}$. Then, $\operatorname{rank} A = r'$ if and only if every $r' + 1 \times r' + 1$ minor of A is zero, and there exists an $r' \times r'$ minor of A that is non-zero.

Proof: See [14, Section 0.4]. ∎

We state and prove a consequence of Lemma 3.1, which we will require for our main result.

Lemma 3.2: Let $A \in \mathbb{R}^{n \times m}$. Then, $\operatorname{rank} A < r'$ if and only if every $r' \times r'$ minor of A is zero.

Proof: First, we show that if $\operatorname{rank} A < r'$, then every $r' \times r'$ minor of A is zero. Let $\operatorname{rank} A = r'' < r'$. Then, by Lemma 3.1, every $r'' + 1 \times r'' + 1$ minor of A is zero. Furthermore, by the Laplace expansion for the determinant [14], for all $r''' \geq r'' + 1$, every $r''' \times r'''$ minor of A is zero. Specifically, since $r' \geq r'' + 1$, every $r' \times r'$ minor of A is zero. Second, we show that if $\operatorname{rank} A \geq r'$, then there exists a nonzero $r' \times r'$ minor of A. Let $\operatorname{rank} A = r'' \geq r'$. By Lemma 3.1 there exists an $r'' \times r''$ nonzero minor of A. It follows from the Laplace expansion for the determinant [14] that for all $r''' \leq r''$ there exists an $r''' \times r'''$ nonzero minor of A. Specifically, there exists an $r' \times r'$ nonzero minor of A. ∎

We next state the main result of this work, a sufficient condition for stationary global identifiability that is based on a set defined by only polynomial equalities.

Theorem 3.2: Consider a chemical reaction network $\mathcal{R}$. If the ideal

$$\mathcal{I} = \big\langle y_j^2 k_j - 1\ \forall j \in \{1, \ldots, r\},\ A_q(\boldsymbol{x}, P)\boldsymbol{k}\ \forall q \in \{1, \ldots r\}, \\ M_i^{(r-1)\times(r-1)}(\boldsymbol{x}, P)\ \forall i \in \{1, \ldots, m\}\big\rangle \tag{11}$$

contains -1, then $\mathcal{R}$ is stationary globally identifiable over $\mathbb{R}^r_{>0}$. Here, $A_q(\boldsymbol{x}, P)$ is the q^{th} row of $A(\boldsymbol{x}, P)$ and $M_i^{(r-1)\times(r-1)}(\boldsymbol{x}, P)$ is all of the size $(r-1) \times (r-1)$ minors of $A(\boldsymbol{x}, P)$, indexed by $i = 1, \ldots, m$.

Proof: Let

$$\bar{V} = \big\{(\boldsymbol{x}, P, \boldsymbol{k}, \boldsymbol{y}) \in (\mathbb{R}^n, \mathbb{S}^{n\times n}, \mathbb{R}^r, \mathbb{R}^r)\big| \\ 0 = A(\boldsymbol{x}, P)\boldsymbol{k}, \\ 0 = M_i^{(r-1)\times(r-1)}(\boldsymbol{x}, P)\ \forall i \in \{1, \dots, m\}, \\ 0 = y_j^2 k_j - 1\ \forall j \in \{1, \dots, r\}\big\}.$$

Recall V defined in (10). We first show that $V = \emptyset$ if and only if $\bar{V} = \emptyset$. First, suppose $V \neq \emptyset$. Then, there exists $(\boldsymbol{x}, P, \boldsymbol{k}) \in V$. It follows that $0 = A(\boldsymbol{x}, P)\boldsymbol{k}$. Let $\boldsymbol{y}$ be such that $y_j = \sqrt{1/k_j}$. Therefore, for all j, $y_j^2 k_j - 1 = 0$. By Lemma 3.2, $\text{rank}(A(\boldsymbol{x}, P)) < r - 1$ guarantees that $0 = M_i^{(r-1)\times(r-1)}(\boldsymbol{x}, P)$ for all $i = 1, \dots, m$, and hence $(\boldsymbol{x}, P, \boldsymbol{k}, \boldsymbol{y}) \in \bar{V}$. Now suppose that $\bar{V} \neq \emptyset$. Then, there exists $(\boldsymbol{x}, P, \boldsymbol{k}, \boldsymbol{y}) \in \bar{V}$. It follows that $0 = A(\boldsymbol{x}, P)\boldsymbol{k}$. Then, we have that $0 = M_i^{(r-1)\times(r-1)}(\boldsymbol{x}, P)$ for all $i = 1, \dots, m$, and hence by Lemma 3.2 it is true that $\text{rank}\, A(\boldsymbol{x}, P) < r-1$. Therefore $(\boldsymbol{x}, P, \boldsymbol{k}) \in V$, and hence $V \neq \emptyset$.

To complete the proof, observe that $\bar{V}$ is the variety of $\mathcal{I}$ defined by (11). If $-1 \in \mathcal{I}$ then by Theorem 2.1, we have that

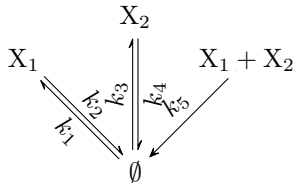
$$\bar{V}_C = \big\{(\boldsymbol{x}, P, \boldsymbol{k}, \boldsymbol{y}) \in (\mathbb{C}^n, \mathbb{C}^{n\times n}, \mathbb{C}^r, \mathbb{R}^r)\big| \\ 0 = A(\boldsymbol{x}, P)\boldsymbol{k}, \\ 0 = M_i^{(r-1)\times(r-1)}(\boldsymbol{x}, P)\ \forall i \in \{1, \dots, m\}, \\ 0 = y_j^2 k_j - 1\ \forall j \in \{1, \dots, r\}\big\} = \emptyset.$$

This immediately implies that $\bar{V} = \emptyset$, and hence by our above argument that $V = \emptyset$. Therefore by Theorem 3.1 $\mathcal{R}$ is stationary globally identifiable over $\mathbb{R}^r_{>0}$. ■

IV. EXAMPLES

In this section, we first present four applications of the tool developed in Section III to certify that a chemical reaction network is stationary globally identifiable.

Example 2 (Mutual degradation): Consider a chemical reaction network $\mathcal{R}_2$ consisting of two species X_1 and X_2 mutually degrading as shown in reactions (12):



(12)

X_1 and X_2 are produced and decay with reaction rate constants k_1 through k_4, and additionally X_1 and X_2 mutually degrade through the reaction $\mathrm{X}_1 + \mathrm{X}_2 \xrightarrow{k_5} \emptyset$. Such a chemical reaction network is an example of a so called *antithetic* motif, and can be used to construct approximate biomolecular realizations of integral controllers [2], [8], [15], [24].

However, the control implemented by the antithetic motif will only have a true integral control term when $k_2 = k_4 = 0$. When $k_2 > 0$ and $k_4 > 0$, a closed loop system constructed using X_1 and X_2 as the controller species will have a nonzero steady state error, because in that case the antithetic motif does not encode a perfect integrator, it instead encodes a "leaky" integrator [24]. Specifically, in [24], it was shown that almost perfect adaptation could be reached as the production rates (parameterized by k_1 and k_3) and the mutual degradation rate (parameterized by k_5) all become much faster than the decay rates (parameterized by k_2 and k_4). This can be used to determine a heuristic to compare two possible biological implementations of $\mathcal{R}_2$ that have parameters $\boldsymbol{k}^A$ and $\boldsymbol{k}^B$, respectively. Specifically, we compute the dimensionless quantities

$$\sigma_1\left(\boldsymbol{k}^A, \boldsymbol{k}^B\right) = \frac{k_2^B k_5^A}{k_5^B k_2^A}, \qquad \sigma_2\left(\boldsymbol{k}^A, \boldsymbol{k}^B\right) = \frac{k_2^B k_1^A}{k_1^B k_2^A}, \\ \sigma_3\left(\boldsymbol{k}^A, \boldsymbol{k}^B\right) = \frac{k_4^B k_5^A}{k_5^B k_4^A}, \qquad \sigma_4\left(\boldsymbol{k}^A, \boldsymbol{k}^B\right) = \frac{k_4^B k_3^A}{k_3^B k_4^A}.$$

If $\sigma_i(\boldsymbol{k}^A, \boldsymbol{k}^B) << 1$ for $i \in \{1, 2, 3, 4\}$, then $\boldsymbol{k}^B$ is expected to perform better than $\boldsymbol{k}^A$.

Here, we study whether a simple experiment that measures only the stationary distribution of (x_1, x_2) with parameters $\boldsymbol{k}^A$ and $\boldsymbol{k}^B$ can be used to estimate $\sigma_i\left(\boldsymbol{k}^A, \boldsymbol{k}^B\right)$, for $i \in \{1, 2, 3, 4\}$. Since stationary global identifiability of $\mathcal{R}_2$ is sufficient to make $\boldsymbol{k}$ identifiable up to a scaling factor, and since for all $\alpha^A, \alpha^B > 0$ we have $\sigma_i\left(\alpha^A \boldsymbol{k}^A, \alpha^B \boldsymbol{k}^B\right) = \sigma_i\left(\boldsymbol{k}^A, \boldsymbol{k}^B\right)$ for $i \in \{1, 2, 3, 4\}$, it is sufficient to check for global identifiability to address this question.

For $\mathcal{R}_2$ we have that

$$\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{k}) = \begin{bmatrix} k_1 - k_2 x_1 - k_5 x_1 x_2 \\ k_3 - k_4 x_2 - k_5 x_1 x_2 \end{bmatrix}$$

and

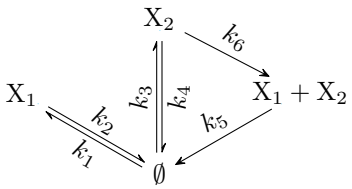
$$\Gamma(\boldsymbol{x}; \boldsymbol{k})\Gamma(\boldsymbol{x}; \boldsymbol{k})^T = \\ \begin{bmatrix} k_1 + k_2 x_1 + k_5 x_1 x_2 & k_5 x_1 x_2 \\ k_5 x_1 x_2 & k_3 + k_4 x_2 + k_5 x_1 x_2 \end{bmatrix}.$$

Therefore, writing (2) in the form (8) yields

$$A(\boldsymbol{x}, P) = \\ \begin{bmatrix} 1 & -x_1 & 0 & 0 & -x_1 x_2 \\ 0 & 0 & 1 & -x_2 & -x_1 x_2 \\ 1 & x_1 - 2p_{11} & 0 & 0 & x_1 x_2 - 2p_{12} x_1 - 2p_{11} x_2 \\ 0 & -p_{12} & 0 & -p_{12} & x_1 x_2 - p_{12} x_1 - p_{12} x_2 \\ & & & & \quad - p_{22} x_1 - p_{11} x_2 \\ 0 & 0 & 1 & x_2 - 2p_{22} & x_1 x_2 - 2p_{22} x_1 - 2p_{12} x_2 \end{bmatrix}. \quad (13)$$

The system $\dot{\boldsymbol{x}} = f(\boldsymbol{x}; \boldsymbol{k})$ has a unique, globally asymptotically stable, equilibrium point in $\mathbb{R}^2_{>0}$ for all $\boldsymbol{k} > 0$ [7]. To prove that $\mathcal{R}_2$ is stationary globally identifiable in the sense of Definition 3.1 we apply Theorem 3.2. In order to use Theorem 3.2, we need to show that $-1 \in \mathcal{I}$, where $\mathcal{I}$ is the ideal defined in (11) with A in (13). To do this, we compute the reduced Gröbner basis $\mathcal{G}$ of $\mathcal{I}$ using the `gbasis` command in Macaulay2 [12] and find that $\mathcal{G} = \{1\}$. Therefore, by Theorem A.1, -1 is in the ideal (11). Hence, by Theorem 3.2, $\mathcal{R}_2$ is stationary globally identifiable over $\mathbb{R}^5_{>0}$.

Example 3 (A nonidentifiable system): We now consider the chemical reaction network $\mathcal{R}_3$:



Chemical reaction network $\mathcal{R}_3$ is the same as $\mathcal{R}_2$, except for the addition of enzymatic production of X_1 by X_2 with reaction rate constant k_6. For $\mathcal{R}_3$, we have from the definition of $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{k})$ in (1) that

$$\boldsymbol{f}(\boldsymbol{x};\boldsymbol{k}) = \begin{bmatrix} k_1 - k_2x_1 - k_5x_1x_2 + k_6x_2 \\ k_3 - k_4x_2 - k_5x_1x_2 \end{bmatrix} \tag{14}$$

and from (3) that

$$\Gamma(\boldsymbol{x};\boldsymbol{k})\Gamma(\boldsymbol{x};\boldsymbol{k})^T = \begin{bmatrix} k_1 + k_2x_1 + k_6x_2 + k_5x_1x_2 & k_5x_1x_2 \\ k_5x_1x_2 & k_3 + k_4x_2 + k_5x_1x_2 \end{bmatrix}.$$

Therefore we have that $A(\boldsymbol{x},P)$ in (8) is given by

$$A(\boldsymbol{x},P)\boldsymbol{k} = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \tag{15}$$

where

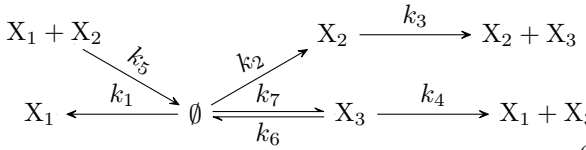
$$A_1 = \begin{bmatrix} 1 & -x_1 & 0 & 0 \\ 0 & 0 & 1 & -x_2 \\ 1 & x_1 - 2p_{11} & 0 & 0 \\ 0 & -p_{12} & 0 & -p_{12} \\ 0 & 0 & 1 & x_2 - 2p_{22} \end{bmatrix}$$

and

$$A_2 = \begin{bmatrix} -x_1x_2 & x_2 \\ -x_1x_2 & 0 \\ x_1x_2 - 2p_{12}x_1 - 2p_{11}x_2 & 2p_{12} + x_2 \\ x_1x_2 - p_{12}x_1 - p_{12}x_2 - p_{22}x_1 - p_{11}x_2 & p_{22} \\ x_1x_2 - 2p_{22}x_1 - 2p_{12}x_2 & 0 \end{bmatrix}.$$

It can be shown that for all $\boldsymbol{k} \in \mathbb{R}^6_{>0}$, (14) has a unique equilibrium point in $\mathbb{R}^2_{\geq 0}$, and that equilibrium point is globally asymptotically stable, see Appendix B. We now show that $\mathcal{R}_3$ is not stationary globally identifiable over $\mathbb{R}^6_{>0}$. First, we note that the lack of identifiability is non-trivial, since in this case $A(\boldsymbol{x},P) \in \mathbb{R}^{5\times 6}$, and so in principal $A(\boldsymbol{x},P)$ could have rank $r-1=5$ for all $(\boldsymbol{x},P)$ such that there exists $\boldsymbol{k}>0$ satisfying $A(\boldsymbol{x},P)\boldsymbol{k}$, which would imply that $\mathcal{R}_3$ is stationary globally identifiable. When $\boldsymbol{k} = \begin{bmatrix}10 & 1 & 10 & 1 & 1 & 10\end{bmatrix}^T$, the solution to $0 = A(\boldsymbol{x},P)\boldsymbol{k}$ is $\boldsymbol{x}^* = \begin{bmatrix}10 & 10/11\end{bmatrix}^T$ and $P^* = \operatorname{diag}\begin{bmatrix}10 & 10/11\end{bmatrix}$. Evaluating the rank of A in (15) with these values of $\boldsymbol{x}$ and P gives $\operatorname{rank} A = 4 < r-1$ and so by Theorem 3.1, $\mathcal{R}_3$ is not stationary globally identifiable over $\mathcal{R}^6_{>0}$.

Example 4 (Antithetic feedback loop): We now consider chemical reaction network $\mathcal{R}_4$, which implements an antithetic feedback loop [2], [8], [15]. The chemical reaction network $\mathcal{R}_4$, which is a model of the system shown in Figure 1(a)(ii) of [24], is given by the reactions shown in (16):



(16)

The controller is implemented by the two species X_1 and X_2, which are each produced at rates k_1 and k_2, and mutually degrade with reaction rate constant k_5. The regulated species X_3 is enzymatically produced from X_2 with reaction rate constant k_3, and the loop is closed by the enzymatic production of X_1 from X_3 with reaction rate constant k_4. The production and decay rate constants of X_3 are k_7 and k_6, respectively. The equilibrium value of x_3 is insensitive to k_3, k_6 and k_7 [8]. We investigate the identifiability of $\mathcal{R}_4$. Due to space constraints we do not explicitly give $A(\boldsymbol{x},P)$ for $\mathcal{R}_4$. However, $A(\boldsymbol{x},P)$ is constructed analogously to the cases of $\mathcal{R}_2$ and $\mathcal{R}_3$ using (1) and (3). It can be shown that $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x};\boldsymbol{k})$ has a unique, locally asymptotically stable equilibrium point in $\mathbb{R}^3_{\geq 0}$ [24]. The global stability of this system is an open question [1], here we simply use the solution to (4) as a model for the stationary distribution of $\mathcal{R}_4$. We use Theorem 3.2 to prove that $\mathcal{R}_4$ is stationary globally identifiable by showing that $-1 \in \mathcal{I}$, where $\mathcal{I}$ is the ideal defined in (11). To do this we compute the reduced Gröbner basis $\mathcal{G}$ of $\mathcal{I}$ using the `gbasis` command in Macaulay2 [12] and find that $\mathcal{G} = \{1\}$. Therefore, by Theorem A.1, -1 is in the ideal (11). Hence, by Theorem 3.2, $\mathcal{R}_4$ is stationary globally identifiable over $\mathbb{R}^7_{>0}$.

V. CONCLUSION

In this work, we study identifiability of chemical reaction networks from stationary distributions. For LNA models, we characterize identifiability by an algebraic condition and check this condition algorithmically by computing the reduced Gröbner basis of a particular ideal. We demonstrate our proposed method by applying it to three different biomolecular systems of practical interest. Our results can be used to determine whether a simple experimental setup wherein measurements are taken from a population of cells grown to steady state is a viable way to estimate the parameters of a biomolecular circuit. Future work includes extending these results to chemical reaction networks modeled by the chemical master equation instead of the LNA, as well as to situations where measurements of the distribution are available at a sequence of times, and/or measurements are only available for a subset of species.

VI. ACKNOWLEDGMENTS

The authors thank Eduardo D. Sontag for several helpful discussions.

APPENDIX

A. Gröbner Bases

We define a monomial as a polynomial $p \in \mathbb{Q}[\boldsymbol{z}]$ that can be written as $p = \prod_{i=1}^N z_i^{\alpha_i}$ for some $N \geq 0$ and $\alpha_1, \alpha_2, \ldots, \alpha_N \in \mathbb{N}$. We let "$\prec$" be a total ordering on the set of monomials in $\mathbb{Q}[\boldsymbol{z}]$ that additionally satisfies

(i) $1 \prec p$ for any nonconstant monomial $p \in \mathbb{Q}[\boldsymbol{z}]$ and
(ii) $\prod_{i=1}^N x_i^{\alpha_i} \prec \prod_{i=1}^N z_i^{\beta_i}$ implies that $\prod_{i=1}^N z_i^{\alpha_i+\gamma_i} \prec \prod_{i=1}^N z_i^{\beta_i+\gamma_i}$ for all $\alpha_1, \ldots, \alpha_N, \beta_1, \ldots, \beta_N, \gamma_1, \ldots, \gamma_N \in \mathbb{N}$

Such a total ordering $\prec$ is called a term order. Let $p \in \mathbb{Q}[\boldsymbol{z}]$. Then, we denote by $\mathrm{in}_{\prec}(p)$ the largest monomial with respect to $\prec$ that appears in p. Suppose that $\mathcal{I} = \langle \mathcal{P} \rangle$. We then have that $\mathcal{G}$ is a Gröbner basis of $\mathcal{I}$ if it is a finite subset of $\mathcal{I}$ that satisfies $\langle \mathrm{in}_{\prec}(p) | p \in \mathcal{I} \rangle = \langle \mathrm{in}_{\prec}(g) | g \in \mathcal{G} \rangle$. $\mathcal{G}$ is a reduced Gröbner basis of $\mathcal{I}$ if additionally

(i) the coefficient of the largest monomial in g with respect to $\prec$ is 1 for each $g \in \mathcal{G}$ and

(ii) for all $g \in \mathcal{G}$, $\langle \mathrm{in}_{\prec}(g') | g' \in \mathcal{G} \setminus \{g\} \rangle$ does not contain any monomial term of g.

The following theorem relates the reduced Gröbner basis of an ideal to the condition $-1 \in \langle \mathcal{P} \rangle$.

Theorem A.1: Let $\mathcal{P} \subseteq \mathbb{Q}[\boldsymbol{z}]$. Let $\mathcal{G}$ be the reduced Gröbner basis of $\langle \mathcal{P} \rangle$. We have that $-1 \in \langle \mathcal{P} \rangle$ if and only if $\mathcal{G} = \{1\}$.

Proof: See e.g. [11], [28]. ∎

Theorem A.1 is convenient for computational purposes since by computing the reduced Gröbner basis of an ideal and then invoking Theorems A.1 and 2.1 we can immediately see if the associated set of polynomial equations has a solution.

B. Stability of Example 3

To show that (14) has a unique equilibrium point in $\mathbb{R}^2_{\geq 0}$ we note that all solutions of $0 = \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{k})$ are given by

$$x_2^* = \frac{-(k_1 - k_3 + \frac{k_2 k_4}{k_5}) \pm \sqrt{(k_1 - k_3 + \frac{k_2 k_4}{k_5})^2 + 4(k_4 + k_6)\frac{k_2 k_4}{k_5}}}{2(k_4 + k_6)},$$

$$x_1^* = \frac{k_1 + k_6 x_2^*}{k_2 + k_5 x_2^*},$$

of which exactly one solution (x_1^*, x_2^*) is nonnegative. To see that (x_1^*, x_2^*) is globally asymptotically stable, observe that $\mathbb{R}^2_{\geq 0}$ is positively invariant and let $W(\boldsymbol{x}) = x_1 + (1 + \frac{k_6}{k_4}) x_2$. We have $\dot{W}(\boldsymbol{x}) \leq k_1 + (1 + \frac{k_6}{k_4}) k_3 - k_2 x_1 - k_4 x_2$, and so there exists $C > 0$ such that for all $\boldsymbol{x} \in \mathbb{R}^2_{\geq 0}$ such that $W(\boldsymbol{x}) \geq C$, $\dot{W}(\boldsymbol{x}) < 0$, and thus all trajectories eventually enter $\mathcal{W} = \{\boldsymbol{x} \in \mathbb{R}^2_{\geq 0} | W(\boldsymbol{x}) \leq C\}$ and stay there. Furthermore, $\frac{\partial \dot{x}_1}{\partial x_1} + \frac{\partial \dot{x}_2}{\partial x_2} < 0$ for all $\boldsymbol{x} \in \mathbb{R}^2_{\geq 0}$, and thus by Dulac's Criterion [16] there are no periodic orbits in $\mathcal{W}$. The Poincaré-Bendixson Theorem then implies that $\boldsymbol{x}^*$ is globally attracting. It can be verified that $\left.\frac{\partial \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{k})}{\partial x}\right|_{\boldsymbol{x} = \boldsymbol{x}^*}$ is Hurwitz [16]. We conclude that $\boldsymbol{x}^*$ is globally asymptotically stable with respect to $\boldsymbol{x} \in \mathbb{R}^2_{\geq 0}$.

References

[1] Deepak K Agrawal, Ryan Marshall, M Ali Al-Radhawi, Vincent Noireaux, and Eduardo D Sontag. Some remarks on robust gene regulation in a biomolecular integral controller. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2820–2825. IEEE, 2019.

[2] Stephanie K. Aoki, Gabriele Lillacci, Ankit Gupta, Armin Baumschlager, David Schweingruber, and Mustafa Khammash. A universal biomolecular integral feedback controller for robust perfect adaptation. *Nature*, 570(7762):533–537, 2019.

[3] S. Audoly, G. Bellu, L. D'Angio, M. P. Saccomani, and C. Cobelli. Global identifiability of nonlinear models of biological systems. *IEEE Transactions on Biomedical Engineering*, 48(1):55–65, 2001.

[4] Michael Backenköhler, Luca Bortolussi, and Verena Wolf. Moment-based parameter estimation for stochastic reaction networks in equilibrium. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(4):1180–1192, 2017.

[5] R. Bellman and K. J. Åström. On structural identifiability. *Mathematical Biosciences*, 7(3):329–339, 1970.

[6] Giuseppina Bellu, Maria Pia Saccomani, Stefania Audoly, and Leontina D'Angiò. Daisy: A new software tool to test global identifiability of biological and physiological systems. *Computer Methods and Programs in Biomedicine*, 88(1):52–61, 2007.

[7] Franco Blanchini and Elisa Franco. Structurally robust biological networks. *BMC Systems Biology*, 5(1):74, 2011.

[8] Corentin Briat, Ankit Gupta, and Mustafa Khammash. Antithetic integral feedback ensures robust perfect adaptation in noisy biomolecular networks. *Cell Systems*, 2(1):15–26, 2016.

[9] Oana-Teodora Chis, Julio R. Banga, and Eva Balsa-Canto. Structural identifiability of systems biology models: A critical comparison of methods. *PLOS ONE*, 6(11):e27755–, 11 2011.

[10] Eugenio Cinquemani. Identifiability and reconstruction of biochemical reaction networks from population snapshot data. *Processes*, 6(9):136, 2018.

[11] David Cox, John Little, and Donal OShea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer Science & Business Media, 2013.

[12] Daniel R. Grayson and Michael E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at `http://www.math.uiuc.edu/Macaulay2/`.

[13] Ankit Gupta, Mustafa Khammash, and Guido Sanguinetti. Bayesian parameter estimation for stochastic reaction networks from steady-state observations. In *International Conference on Computational Methods in Systems Biology*, pages 342–346. Springer, 2019.

[14] Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[15] Hsin-Ho Huang, Yili Qian, and Domitilla Del Vecchio. A quasi-integral controller for adaptation of genetic modules to variable ribosome demand. *Nature Communications*, 9(1):5415, 2018.

[16] Hassan K Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2002.

[17] Thomas G Kurtz. Strong approximation theorems for density dependent markov chains. *Stochastic Processes and their Applications*, 6(3):223–240, 1978.

[18] Gabriele Lillacci and Mustafa Khammash. The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics*, 29(18):2311–2319, 6/26/2021 2013.

[19] Lennart Ljung and Torkel Glad. On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2):265–276, 1994.

[20] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[21] Brian Munsky, Guoliang Li, Zachary R. Fox, Douglas P. Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences*, 115(29):7533–7538, 2018.

[22] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular systems biology*, 5(1):318, 2009.

[23] Kaan Öcal, Ramon Grima, and Guido Sanguinetti. Parameter estimation for biochemical reaction networks using wasserstein distances. *Journal of Physics A: Mathematical and Theoretical*, 53(3):034002, Dec 2019.

[24] Yili Qian and Domitilla Del Vecchio. Realizing 'integral control' in living cells: how to overcome leaky integration due to dilution? *Journal of The Royal Society Interface*, 15(139), 2018.

[25] Eduardo D Sontag. On the observability of polynomial systems, i: Finite-time problems. *SIAM Journal on Control and Optimization*, 17(1):139–151, 1979.

[26] Eduardo D Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Science & Business Media, 2013.

[27] Ralf Steuer, Jürgen Kurths, Oliver Fiehn, and Wolfram Weckwerth. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19(8):1019–1026, 2003.

[28] Bernd Sturmfels. What is... a grobner basis? *Notices-American Mathematical Society*, 52(10):1199, 2005.

[29] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[30] Alejandro F. Villaverde, Antonio Barreiro, and Antonis Papachristodoulou. Structural identifiability of dynamic systems biology models. *PLOS Computational Biology*, 12(10):e1005153–, 10 2016.